Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.5, No.15, 2014

The Effect of Performance Based Assessment on Language Accuracy of Tenth Grade English Language Students at Mafraq Borough Directorate of Education

Rania Shaddad Qutaishat & Ahmad Musa Bataineh
rania shaddad qutesha& Ahmad Musa Bataineh *
* Raniashaddad@gmail.com

Abstract
The purpose of this study is to investigate the effect of using performance-based assessment on tenth grade students' language accuracy achievement in Mafraq Borough Directorate- Jordan. Equivalent pre/post-test two group design was used. through a pre/post-test comprehension test with grammar components was constructed to measure students' achievement in language accuracy. The sample of this study consisted of 128 tenth grade students; sixty-four were male students from The First Secondary School for Boys, and sixty-four were female students from Princess Alia Bent Al-Hussein School for Girls during the first semester of the academic year 2009/2010. The subjects of the study were distributed among four groups (a female experimental group and control group, and a male experimental group and control group). The experimental groups were evaluated using performance based Assessment, while the control groups were evaluated via traditional pencil and paper test. The subjects were (32) male students for the experimental group and (32) male students for the control group, while the female students for the experimental and control group were (32) and (32) respectively. Those subjects were distributed among four purposefully selected sections in two schools that were also selected purposefully. The findings of the study indicated that there were statistically significant differences in the post- test between the control group and the experimental group in favor of the experimental group, and there was a statistically significant difference in the students' achievement due to gender in favor of females. There was statistically significant difference due to the interaction between gender and group in favor of female students in the experimental group.
Keywords: Performance Based Assessment. Language Accuracy. 10th grade students. Mafraq. Jordan

1. Introduction & Literature review

Assessment is central to teaching and learning. The assessment information is needed to make informed decisions regarding students' learning abilities, their placement in appropriate levels and their achievement.
According to Sadler (2005), assessment refers to the making of evaluation on students' overall performance and generating assumptions regarding their learning and production education-wise, which include the quality or achievement in tasks such as tests, projects, reports and examinations. In the other hand, the success of any assessment is depending on the effective selection and use of appropriate procedures as well as on the proper interpretation of students' performance. The correlations between students' preferences and assessment perceptions in their findings were not significant due to the existence of a distinction between students' preferences and their perceptions.
Today, a common method advocated to improve students' achievement is the use of formative assessments, both to improve the pedagogical practices of teachers and to provide specific instructional support for lower performing students (Dunn and Mulvenon, 2009). In fact, the formative assessment methods of assessing students take into account variation in students' needs, interests and learning styles; and they attempt to integrate assessment and learning activities. In the integral process of learning and instruction only high quality assessment can facilitate high-quality learning. Mueller (2005) observes that while researchers in higher education have proposed a series of changes such as alternative assessment to replace traditional assessment, these proposals have yet to be implemented in many institutions.
Many researchers found that improved formative or classroom assessment practices helped low achievers more than other students. This revealing finding has direct implications for school systems that want to bridge the achievement gap. To make improvements, however, teachers must be provided with the assessment tools that they need to increase the achievement of English language learners. New understandings of the learning process indicate that assessment and learning are intimately linked. These new understandings of learning need to be applied to classroom-based assessment practices (Marzano, Pickering, & McTighe, 1993). Among these practices, performance-based assessment appears to hold promise for improving the educational attainment of English language learners.

College of Education, Department of Curriculum and Instruction, Yarmouk University.

The term "performance assessment" is not new. According to Ryans and Frederiksen (1951), performance tests had been used extensively for measuring general level of ability of individuals who suffered from language deficiencies. In this sense, a performance test was more or less synonymous with "nonverbal test" and was used primarily to distinguish this type of measurement from that requiring linguistic ability. Furthermore, Ryans and Frederisken distinguished between "knowledge or information" measured by verbal tests and "skills", such as cooking, driving, teaching, and etc., measured by performance tests. All these forms of performance test are used to measure directly the skills of interest insofar as the skills are readily apparent in the performance or products that are generated. Thus, the core of a performance task is an instance of real-life application.

Following this tradition, current discourse on performance-based assessment explicitly extends its use in assessing workplace performance to the assessment of achievement in school subject matters. For example, Linn, Baker, and Dunbar (1991) described performance-based assessment as one that "closely emulates the mental tasks of life and the everyday workplace seems natural" (p. 2). Shavelson (1994) referred to performance-based assessment as "concrete, well contextualized tasks or activities that are sampled from a subject matter and that often call for an integration of conceptual understanding with performance skills ... the response required by the assessment is similar to the production expected of someone practicing the subject matter domain" (p. 235). Gitomer (1993) defined a performance task as "one that simultaneously requires the muse of knowledge, skills, and values that are recognized as important in a domain of study and is qualitatively consistent with tasks that members of discipline-based communities might conceivably engage in. Assessment entails judgments and reports of the quality of performance by community members" (p. 244).

Description

Performance assessment strategies are composed of three distinct parts: a performance *task*; a *format* in which the student responds; and a predetermined *scoring system*. Tasks are assignments designed to assess a student's ability to manipulate equipment (laboratory equipment, computers, documents, etc.) for a given purpose. Students can either complete the task in front of a panel of judges or use a written response sheet. The student is then scored by comparing the performance against a set of written criteria. When used with students with highly varying abilities, performance tasks can take maximum advantage of judging student abilities by using tasks with multiple correct solutions. Students are graded on the process of problem solving using a rating scale based on explicit standards.

Using performance-based assessment to promote learning

Classroom-based assessments may be of two broad types: selected-response and constructed-response formats. Selected-response formats provide response items for students to choose from (such as multiple-choice, true-false, and matching items). Constructed response formats, on the other hand, ask students to develop a response, create a product, or conduct a demonstration (Feuer & Fulton, 1993; Frisby, 2001; Herman, Aschbacher, & Winters, 1992; McTighe & Ferrara, 1998). These types of

assessments allow more than one correct answer to a problem and typically involve higher-order thinking skills. Performance-based assessment (PBA), which uses a constructed-response format, has as its primary purpose the improvement of learning. Performance based assessment links assessment to instruction through the use of meaningful and engaging tasks. Performance tasks may also call for integration of language and content-area skills. Authentic assessment, a type of PBA, promotes application of knowledge and skills in situations that closely resemble those of the real world (Frisby, 2001; McTighe & Ferrara, 1998; Wiggins, 1998). Authentic assessments are potentially more motivating than other types because they engage students in realistic uses of language and content area concepts. Authentic assessment and other types of PBA can be used in the service of education to promote transfer or generalizability of learning from facts and procedures to applications in meaningful contexts. A large range and number of tasks are needed over time, however, to ensure the generalizability of performance-based assessment. Can performance-based assessments be used to monitor and support the learning of English language learners? A number of factors make performance-based assessment more appropriate for English language learners than traditional testing formats (Frisby, 2001; Hamayan & Damico, 1991; O'Malley & Pierce, 1996).

Performances allow students to demonstrate application of their knowledge and skills under the direct observation of the teacher. Students may engage in tasks that are useful outside of school, such as asking for directions by telephone, demonstrating a process, or arguing a position. All of these can demand high levels of language skills. Examples of performance tasks include oral reports, skits and role plays, demonstrations, and debates.

Process-oriented assessments provide insight into students' thinking, reasoning and motivation. They can provide diagnostic information on how well students use learning strategies and may lead to independent learning when students are asked to reflect on their learning and set goals to improve it. Some examples of process-oriented assessments are think-aloud, self assessment checklists or surveys, learning logs and individual

or pair conferences. Products, performances and process oriented assessments can all be used to generate rich information on English language learners' ability to transfer learning and meet state and local standards. Two features of performance-based assessment help in supporting the development of mental habits that lead to independent learning. The first is referred to as *visible criteria*. A fundamental tenet of performance-based assessment is the sharing of standards and making the criteria for evaluation, visible to students. The second key element of performance- based assessment is **self-assessment,** which is essential for teaching students how to manage their study habits, use learning strategies, and reflect on progress toward learning goals. The goal of self-assessment is to produce students who can learn independently of the teacher and become lifelong learners. To accomplish this, teachers need to provide students with specific feedback, opportunities to give and receive feedback from peers, and time to set learning goals ERIC/CLL (News Bulletin, fall 2002 • page 3.)

Authentic assessment emphasizes the practical application of tasks in real-world settings. Mueller (2005) defines authentic assessments as direct measures of students' acquired knowledge and skills through formal education to perform authentic tasks. The realistic contexts can make problems more engaging for students and help the teachers evaluate whether a student who can solve a problem in one context can transfer the skills to a similar setting. Besides that, research has conclusively demonstrated that the use of formative assessment facilitates improvement in instructional practices, identifies "gaps" in the curriculum and contributes to increased student performance (Dunn and Mulvenon, 2009). Hence, to perform these authentic tasks, students need to construct their own meaning to the world through the application of previously acquired information from classroom teaching and learning (Airasian, 2005; Linn and Miller, 2005).

1.1 Statement of Purpose

The ability to write correct sentences using correct grammar and vocabulary is one of the most essential skills EFL learners need to develop through their schooling. However, after years of learning English as a foreign language, students in the Jordanian schools appear to be unable to write properly. This could be ascribed to a number of factors. The way in which English is taught is believed to be a decisive factor. Although the current English language syllabus, as stated in the English language curriculum and the English language teacher's book, is based upon the communicative approach to foreign language teaching. The practice of teaching language seems to be carried out in a traditional way. Teachers enter the class, give their students a pencil and paper test, and ask them to answer it. Then, teachers evaluate the students, commenting on mechanics of writing and language elements (grammar, spelling, punctuation, handwriting, and vocabulary) of students' answers. This resulted in students' failure to express them selves properly in writing. A great number of Jordanian students do not appear to write English very well. Furthermore, teachers complain that students are barely capable to produce grammatically correct sentences. On the other hand, teachers suffer a lot when they are assessing and evaluating their students, they think that they are unfair in testing and evaluating, they know how to teach but they do not know how to test. These are some of the complaints one frequently hears about when he means Jordanian students studying English. Therefore, assessing students should be made much more effective.

1.2 Study objectives and Questions

The purpose of this study is to determine whether the use of performance-based assessment is effective in evaluating students' language accuracy in English language compared to the traditional method used in Jordanian public schools. This study attempts to answer the following questions:

1-Are there any statistically significant differences (α=0, 05) between the groups due to method of evaluation (Performance-based assessment and traditional method)?

2-Are there any statistically significant differences (α=0, 05) between the groups due to gender (male or female)?

3- Are there any statistically significant difference (α=0, 05) in the interaction between groups and the gender (male or female)?

1.3 Significance of the study

The current study on the effect of performance-based assessment on students' language accuracy is expected to serve two goals: To help with integrating new ways of evaluation into English language classes to meet the new century demand, and to find solutions to some of the problems of language accuracy in Jordanian schools. The choice of the topic for this study is motivated by several factors. Firstly, the study responds to the increased demand in the use of performance-based assessment in education to meet the new educational needs. Secondly, the study may motivate other researchers to reconsider the methods of evaluation used nowadays. And finally, the performance-based assessment procedures might be a source of excitement and motivation to Jordanian students in their English language classes

1.4 Study Limitations

This study is limited to the male and female 10th grade students in Mafraq Borough Directorate, and to any other similar samples. Besides, it is limited to the instrument which was developed by the researchers for the sake of the study. The generalization of the study findings will be within the context of study restrictions.

1.5 Definition of terms

Assessment: A process by which a person's knowledge and ability can be measured (Dowing, 1992).

Performance-based assessment: Assessment measures which are performance-based and require the learner to use essential knowledge or to demonstrate the acquisition of targeted outcomes by demonstrating real life tasks or approximations of them. Such examples are writing assignments, experiments, simulations, hands-on activities, model construction, etc. (Burstein, 1991).
Accuracy: The ability to produce correct sentences using correct grammar and vocabulary; or the ability to produce sentences that are semantically, syntactically and socio-culturally well formed.
Traditional assessment: An assessment measure which offers the selection of existing possibilities such as, matching, multiple-choice, completion, true-false, or short answers, etc., for evaluation (Dowing, 1992).
Action Pack: The textbook which is taught from grade one to grade twelve in Jordanian public schools. It is a twelve-level communicative language program, each level of Action Pack contains the following components: A student book, workbook, teachers' book and cassettes (Hains, 2008).

2. Previous studies

Several studies were conducted to investigate the effect of using performance-based assessment on students' language accuracy in English. For example, Clark and Gognet (1985) reported on how they developed and validated a performance-based test of ESL survival skills. Wesche (1987) reported on the development of an integrated skills (reading, writing, listening, & speaking) English for academic purposes performance test called the *The Ontario Test of ESL*, which contained a general academic English section, plus a discipline-related thematic section (examinees could choose either science or social science). McNamara (1990) investigated the effectiveness of using item response theory to develop and validate an English-for-specific-purposes performance test for health professionals. Shameem (1998) studied the relationship between self-reported language proficiency and performance tests developed for the Indo-Fijian immigrant community in Wellington, New Zealand. North and Schneider (1998) report on the use of Rasch analysis to empirically develop and validate scale descriptors for proficiency assessment based on language performance for English, French, and German in a Swiss National Science Research Council project.
Moreover, Related to *writing*, Allaei and Connor (1991) report on developing and using performance tests for assessing ESL writing ability. Two other studies empirically investigated the sorts of writing tasks that are required in American academic degree programs, one by examining the actual writing assignments of students at a single university (Horowitz, 1986), and the other through a wide ranging and carefully designed survey study (Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996). With regard to *listening*, Scott, Stansfield, and Kenyon (1996) investigated the validity of a "summary translation" performance test of Spanish language listening ability. Stansfield, Wu, and van der Heide (2000) reported on the development and validation of a job-relevant listening summary translation performance test for selection and placement of speakers of Minnan who were employees of the U.S. Government. Brindley and Slatyer (2002) investigated the difficulty of tasks in ESL listening assessment based on actual performances of adult ESL learners in Australia. O'Sullivan (2002) studied the validity of tasks on UCLES EFL speaking tests by comparing the results to *a priori* and *a posteriori* analyses of the actual speaking task output.
In terms of *speaking*, Stansfield and Kenyon (1992) report on the development and validation of a simulated oral proficiency interview. Douglas and Selinker (1993) studied the relationship between performances of international teaching assistants on a general speaking test and a test designed to be discipline-specific. Fulcher (1996) studied task design and group oral performance tests from the students' points of view. Kenyon (1998) investigated the validity of tasks on performance-based tests of oral proficiency of German, French, and Spanish students at the high school and college levels. Hill (1998) studied the effectiveness of validating an oral English proficiency test through test-takers reactions to and performance on a test performance test for prospective migrants to Australia. Chalhoub-Deville (2001) examined the task-based validity of three oral assessments (an oral proficiency interview, a contextualized speaking assessment, and a video/oral communication instrument).
With regard to *integrated skills*, in what Bachman (2002, p. 454) referred to as the "most fully conceptualized, operationalized and researched exemplification of this approach of which I am familiar.

3. Study Methodology

3.1 Methodology

It contains a description of the sample of the study, and study tool, and procedures for validity and reliability of the instrument used in the study, also deal with a description of the statistic that will be used in the analysis of data, and extract the results, this study belongs to a type of descriptive research survey aimed to, analysis, and evaluate of the characteristics of a particular group, or a certain position dominated by the recipe selection

3.2 Study Subjects

The subjects of the study consisted of 128 tenth grade students and distributed on four sections, which were selected purposefully. Two female sections were chosen from Princess Alia Bent Al-Hussein School for Girls and the other two sections were chosen from The First Secondary School for Boys. In each school, one section

was assigned as an experimental group and the other as a control group; table (1) shows the distribution of the subjects of the study according to group and gender variables.

Table (1): Distribution of the Subjects of the Study According to Group and Gender Variables

Group	Gender	N	Percentage
Experimental	Male	32	0.25
	Female	32	0.25
Control	Male	32	0.25
	Female	32	0.25

Table (1) shows that the participants in the study were male and female students in both of the experimental and control groups, the number of the female students was equal to male students, which reached (64) for each group distributing on male and female sections (32) for each section.

3. 3 Study Instrument

The researchers developed a test based on the instructional material of the 10th Grade Students' Book to collect the data. The test was prepared by the researcher. She validated it and made it reliable. Both groups; the experimental group as well the control group, were taught by the researcher her self. The subjects in both groups underwent a pre-test to determine their actual level before starting the experiment, and the same test was administered as a post-test at the end of the experiment to assess subjects' achievement. The time interval between the pre-test and the post-test was (16) weeks; a period long enough to minimize the effect of the pre-test on the results and conclusions of the experiment. The test contained three reading passages followed by comprehension questions.

The grammar component consisted of a set of grammar questions. The vocabulary part consisted of two questions, and the writing part consisted of three questions. The questions were Wh-questions, multiple choice questions, sentence completion and matching words with their meanings.

3.4 Instrument Validity

The researchers constructed the test instrument taking into consideration the instructional material. The researcher validated the instrument by consulting two TEFL professors teaching at Al-Albayt University, three supervisors of English working at the Directorate of Education, and two 10th grade teachers of English. The researcher followed the recommendations of the referees and made amendments accordingly. When producing the final version of the test, the remarks and recommendations of these EFL experts were taken into account.

3.5 Instrument reliability

To ensure the test reliability, the researchers followed test/retest technique. The researcher administered it to a pilot sample of (20) subjects out side the study sample in the same city from which the subjects were chosen with a two-week period between the pre-test and the post-test. The reliability of the test was concluded using correlation coefficient and found to be 0.87. The researcher considered this value acceptable for the purposes of the study.

3.6 Instructional material

The instructional material is the tenth grade English textbook entitled "Action Pack 10". It consists of twelve units, but the researcher supposed to cover at least eight units during the application period. Each unit has five components: Warming up, Presentation, Application, Practice and Assessment. In the warming up, the textbook presents the topic through a set of general questions and photos to present the unit. In the presentation stage, the textbook presents new ideas, vocabulary items, reading comprehension text and grammatical aspects as well. In the application stage, the new ideas, the vocabulary items and the grammatical aspects are applied in various situations. In the practice stage, the students are allowed to practice the new ideas, the vocabulary items and the grammatical aspects given in the previous stages. And finally, in the assessment stage, the items are evaluated through a set of exercises. And at the end of each three units, there is a revision part, which revises units through exercises and gives a sample test to cover the units (Haines, 2008).

Furthermore, in the assessment stage, the researchers use the assessment exercises at the end of each unit. They also developed worksheets containing comprehension exercises followed by grammar questions related to the reading comprehension provided.

4. Study Findings

The purpose of this study is to investigate the effect of using performance-based assessment on tenth grade students' language accuracy achievement in Mafraq Borough Directorate- Jordan. The researchers followed the equivalent pre /post test two group designs. Therefore, the means, standard deviations and Two-Way ANOVA analysis of variance were used to analyze data. The results will be displayed based on the questions of the research.

Equivalent test

Means, standard deviation and (Two-Way ANOVA) analysis of variance were used to analyze data on the pre-test. In order to determine the equivalency of the two groups on pre-test, the means and the standard deviation

were computed, as shown in table (2).

Table (2): Means and Standard Deviation of Students' Scores on the Pre-test.

Group	Gender	N	Mean	SD
Experimental	Male	32	21.63	10.97
	Female	32	21.81	7.74
	Total	64	21.72	9.4
Control	Male	32	18.34	8.57
	Female	32	21.63	10.97
	Total	64	20.06	9.5

Table (2) shows that the mean scores and standard deviation of pre-test both by group and by gender were close to each other.

In order to find out if there are statistically significant differences between the two groups before treatment, Two-Way ANOVA analysis of variance was performed. The results are reported in table (3).

Table (3): Results of ANOVA Analysis of Students' Scores on the Pre-test.

Source	Sum of Squares	DF	Mean Square	F	Sig.
Group	105.125	1	105.125	1.166	0.282
Gender	87.781	1	87.781	0.974	0.326
Group*Gender	84.500	1	84.500	0.938	0.335
Error	11175.063	124	90.121		
Corrected total	11452.469	127			

The results of this analysis indicated that there were no statistically significant differences in the pre-test scores between the experimental and control groups. There was also no statistically significant difference in students' scores due to their gender. This indicates that the experimental and the control groups were almost equivalent before treatment. The results displayed based on the questions of the study as follows:

4.1 The findings of the first question: The statistical analysis was used to calculate the means and standard deviations for post data of language accuracy within two groups, control and experimental, as shown in table (4).

Table (4): Means and Standard Deviation of Students' Scores on the Post-Test.

Group	Gender	N	Mean	SD
Experimental	Male	32	33.34	4.47
	Female	32	33.59	5.99
	Total	64	5.2	5.2
Control	Male	32	23.66	4.53
	Female	32	29.09	4.08
	Total	64	26.37	5.0
Total	Male	64	28.5	6.61
	Female	64	31.34	5.56
	Total	128	29.92	6.25

It is obvious that the experimental group got a higher mean score than the control group; this indicates that performance-based method has strong impact on the experimental group.

To find out if there were statistically significant differences between the two groups' scores on the post-test, (Two Way ANOVA) analysis was performed, as presented in table (5).

Table (5):Results of (ANOVA) Analysis of the Students' Scores of the Two Groups on the Post Test.

Source	Sum of Squares	DF	Mean Square	F	Sig.
Group	258.781	1	258.781	11.123	0.001
Gender	1610.281	1	1610.281	69.214	0.000
Group*Gender	215.281	1	215.281	9.253	0.003
Error	2884.875	124	23.265		
Corrected total	4969.219	127			

Table (5) shows that there were statistically significant differences in the post- test between the control group and the experimental group in favor of the experimental group.

This means that the intervention of performance-based method affected students' language accuracy.

4.2 The findings of the second question: As shown in table (5), the results showed that there was a statistically significant difference in the students' achievement due to their gender in favor of female.

4.3 The findings of the third question: As shown in table (5), the results also showed that there was a statistically significant difference (α=0, 05) due to the interaction between gender and group. Post-Hoc Tests (Scheffe) was applied for comparisons between sub groups, tables (6, 7) show that.

Table (6): Post-Hoc Tests (Scheffe) Result for Comparisons Between Sub Groups

Multiple Comparisons. Dependent Variable: Test Result. Scheffe

(1) GROUPS	(J) GROUPS	Mean Differences	Std. Error	Sig.
Male control	female control	-6.2812*	1.33127	.000
	male experimental	-10.2812*	1.33127	.000
	female experimental	-12.0937*	1.33127	.000
Female control	male control	6.2812*	1.33127	.000
	male experimental	-4.0000*	1.33127	.033
	female experimental	-5.8128*	1.33127	.000
Male experimental	male control	10.2812*	1.33127	.000
	female control	4.0000*	1.33127	.033
	female experimental	-1.8125	1.33127	.605
Female experimental	male control	12.0937*	1.33127	.000
	female control	5.8125*	1.33127	.000
	male experimental	1.8125	1.33127	.605

Based on observed means.

- The mean difference is significant at the .05 level.

Table (7): Distribution of the Means for Subset Groups as a Result of Scheffe Test.

GROUPS	N	Subset 1	2	3
Male	32	26.0313		
Female	32		32.3125	
Male experimental	32			36.3125
Female experimental	32			38.1250
Sig.		1.000	1.000	.605

Means for groups in homogeneous subsets are displayed. Based on Type 11 Sum of squares
The error term is Mean Square (Error) = 28.357. a.Uses Harmonic Mean Sample Size= 32.000 b. Alpha= .05.

Tables (6, 7) showed the following:

1- There were significant differences between male and female in control group in favor of female, whenever the means were (26.03), (32.31) respectively.
2- There were no significant differences between male and female in experimental group, whenever the means are (36.31), (38.12) respectively.
3- There were significant differences between male in control group and male in experimental group in favor of male in experimental group, whenever the means were (26.03), (36.31) respectively.
4- There were significant differences between male in control group and female in experimental group in favor of female in experimental, whenever the means were(26.03), (38.12) respectively.
5- There were significant differences between female in control group and male in experimental group in favor of male in experimental, whenever the means were (32.31), (36.31) respectively.
6- There were significant differences between female in control group and female in experimental group in favor of female in experimental, whenever the means were (32.31), (38.12) respectively.

5. Conclusion

The findings of the study indicated that alternative and authentic assessment have more acceptance among students, therefore it is viewed as an alternative to traditional standardized assessment. The assessment methods employed during the semester were a new experience to all students in the experimental group. The assignments gave the teachers and students an insight into their group members' thought processes and the evidence gathered during the process. Many students enjoyed the experience. This study also revealed that the assessment criteria and practice need to be explored further to improve the validity and reliability and therefore fairness of assessment practices. Furthermore, the assessment of students' progress and achievement should be carried out in a manner that does not cause anxiety to the students. The summative form of testing that permeated the traditional curricula would not be fair to students. Hence, the traditional paper-and pencil tests no longer cover the variety of activities and tasks that take place in the classroom. The findings have witnessed a major shift from strictly summative testing tools and procedures to a more humanistic, authentic and informal techniques that stress formative assessment.

6. Discussion

The results of the study revealed that using performance-based assessment has a positive effect on the

achievement of the students, the findings of the study indicated that there were no statistically significant differences in the pre-test scores between the experimental and control groups (F=1.166, Sig.=0.282). There was also no statistically significant difference in students' scores due to their gender (F=0.974, Sig. = 0.326). This indicated that the experimental and the control groups were equivalent before treatment.

After treatment, the experimental group got higher mean scores than the control group which were 33.47 and 26.37 respectively. The study also showed that there was statistically significant difference in a post-test between the control group and the experimental group (F=11.123, Sig. = 0.001) in favor of the experimental group and this means that the using of performance-based assessment is better than using the traditional assessment in enhancing students' language accuracy. It is evident that the experimental groups performed much better on the post-test than the control groups. Thus, it could be concluded that the students who were assessed by using performance-based assessment scored significantly higher in the pos- test than the students who were assessed by traditional assessment at(α=0, 05) . The findings of the study indicated that there was statistically significant difference in the students' achievement due to their gender (F=69.214, Sig. = 0.000).The results were in favor of female students and this was expected because girls are always better than boys in language even in their native language. Furthermore, the results showed that there was statistically significant difference (α=0, 05) due to the interaction between gender and group (F= 9.253, Sig. =0.003).

7. Recommendations

The following are recommendations for research:

If this study is to be replicated to bring further significance, some changes should be made

- Perform the experiment over a longer period of time so that students have adequate time to shake off current habits of traditional assessment and become more familiar with the performance-based assessment.
- Conducting other studies to investigate the effect of performance-based assessment on other language skills such as speaking and listening.
- Conducting other studies on other classes in addition to tenth graders so that more students of different levels may be included to make generalizations more valid.

References

Airasian,W., 2005. Classroom assessment: Concepts and applications. 5th Edn., McGraw-Hill, Boston, ISBN: 0-07-248869-7, pp: 234.

Allaei, K., & Connor, U. (1991). Using performative assessment instruments with ESL student writers. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 227-240). Norwood, NJ: Ablex.

Bachman, F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 239-257.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369-394.

Burstein, L. (1991). Performance Assessment for Accountability Purposes: Taking the Plunge and Assessing the Consequences. Center for Research on Evaluation Standards, and Student Testing. Paper presented at *The Annual Meeting of the American Educational Research Association, Chicago*, IL.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210-228). Harlow, UK: Pearson Education.

Clark, J., & Grognet, G. (1985). Development and validation of a performancebased test of ESL "survival skills." In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 89-110). Ottawa: University of Ottawa Press.

Dowing, M. (1992). True-false, alternate- choice, and multiple- choice items. *Educational Measurement: Issues and Practice*, 11 (3).

Douglas, D., & Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In D. Douglas, & C. Chapelle (eds.), *A new decade of language testing* (pp. 235-56). Alexandria, VA: TESOL.

Dunbar, B., Koretz, M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4(4),* 289-303.

Dunn, E. and S.W. Mulvenon, (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. Pract. Assess. Res. Evaluat., 14: 1 11.

Feuer, J., & K, Fulton, (1993). The many faces of performance assessment. *Phi Delta Kappan, 74* (6), 478.

Frisby, L. (2001) Academic achievement. In L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment* (2nd ed., pp. 541-568). San Francisco: Jossey- Bass.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.
Gitomer, D. H. (1993). Performance assessment and educational measurement. In R.. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 241-263). Hillsdale, NJ: Erlbaum.
Haines, S., (2008). Action Pack, York press, England.
Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. TOEFL Research Report # 54.
Hamayan, V., & Damico, S. (1991). *Limiting bias in the assessment of bilingual students*. Austin, TX: Pro-Ed.
Linn, L. and M. Miller, (2005). *Measurement and Assessment in Teaching*. 9th Edn., Pearson, New Jersey, ISBN: 0-13-127393-0, pp: 250-256.
Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
Hill, K. (1998). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A. J. Kunnan (Ed.), *Validation in Language Assessment* (pp. 209-229). Mahwah, NJ: Lawrence Erlbaum Associates.
Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*(3), 445-462.
Linn, L., Baker, L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher,* 20(8), 15 21.
Marzano, J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Alexandria, VA: Association for Supervision and Curriculum Development.
McNamara, T. (1996). *Measuring second language performance: A new era in language testing.* New York: Longman.
McTighe, J., & Ferrara, S. (1998). *Assessing learning in the classroom.* Washington, DC: National Education Association. No Child Left Behind Act (2001). Part A, Improving Basic Programs operated by Local Education Agencies, Subpart 1, Basic Program Requirements. Section 1001 (3): Statement of purpose.
Mueller, J., (2005). The authentic assessment toolbox: Enhancing students' learning through online faculty development. J. Online Learn. Teach., 1: 1-7. http://jolt.merlot.org/documents/vol1_no1_mueller _001.pdf
North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.
O'Malley, M., & Pierce, V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers.* New York: Longman.
O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.
Ryans, G., & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 455- 494). Washington, DC: American Council on Education.
Sadler, R., 2005. Interpretations of criteria-based assessment and grading in higher education. Assess. Evaluat. Higher Educ., 30: 175-194.
Scott, L., Stansfield, C., & Kenyon, D. (1996). Examining validity in a performance test: The listening summary translation exam (LSTE)—Spanish version. *Language Testing, 13*(1), 83-109.
Shameem, N. (1998). Validating self-reported language proficiency by testing performance in an immigrant community: The Wellington Indo-Fijians. *Language Testing, 15*(1), 86-108.
Shavelson, J. (1994). Guest editor's preface. *International Journal of Educational Research,* 27(3), 235-237.
Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*(2), 129-141.
Stansfield, W., Wu, M., & Van der Heide, M. (2000). A job-relevant listening summary translation exam in Minnan. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 177-200). Cambridge: Cambridge University.
Wesche, M. B. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing, 4*, 28-47.
Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Acknowledgment
The authors would like to express greatest and deepest gratitude to all those who assist us, for their assistance and insightful comments on the drafts of this paper. We would like also to thank all the participants in the present study for their cooperation.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Recent conferences: http://www.iiste.org/conference/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar