Research on Humanities and Social Sciences
ISSN 2224-5766 (Paper) ISSN 2225-0484 (Online)
Vol.8, No.3, 2018

The Effect of Classroom Observation and Teacher-Portfolio Evaluation on the General Language Proficiency Achievement of EFL learners: A Case of Iranian Intermediate Students

Masood Khalili Sabet, PhD
University of Guilan, Rasht, Guilan, Iran

Amir Mahdavi Zafarghandi, PhD
University of Guilan, Rasht, Guilan, Iran

Azadeh Khoshsorur
University of Guilan, Rasht, Guilan, Iran

Abstract
The present study attempts to investigate the effect of teacher evaluation on Iranian intermediate EFL learners' general language proficiency achievement. Teacher evaluation is widely considered to be the important point in any educational system which aims at evaluating and assessing the performance of the teachers. Various techniques and methods have been suggested and practiced in educational contexts. This study employs the two well-known techniques for evaluation of the teachers: classroom observation and teacher portfolio. In doing so, 45 male and female English learners, who were all at true intermediate level in language institute Sama in Rasht, Gulian, Iran were randomly selected. These learners were divided in three groups: two experimental groups and one control group. These three groups received a pre-test before the treatment sessions (the control group did not receive any treatment). Finally the post-test which was the same as the pre-test was administrated. This research was conducted for about six months and a half during which two semesters took place. During treatment sessions each of the teachers of the experimental groups received both techniques. But in order to minimize any irrelevant teacher effect, each teacher for a semester received one treatment. In other words, one teacher dealt with classroom observation and the other teacher received portfolio treatment in the first semester. Similarly, for the second term the operation of the methods changed. As a matter of fact, the first teacher received portfolio and the second teacher was evaluated with classroom observation. Finally, the scores of the students at the post-test phase were calculated through the statistical package of SPSS. The results revealed that teacher-portfolio evaluation and classroom observation had a significant effect on the general language proficiency achievement of the learners and could be beneficial for enhancing the academic achievement of the learners.
Keywords: teacher evaluation, general language proficiency achievement, intermediate level, teacher portfolio, classroom observation, Iranian EFL learners

1. Introduction

Identifying the relationship between teacher evaluation and student achievement has been the core of debate among several experts. Therefore, significant changes in the system of evaluation have been proposed and conducted by various stakeholders during the last decades. According to Ellett & Teddlle(2003), innovations in these fields have been concerned: applying the most well-known theories and instructional models such as constructivism and cooperative learning, and updating current approaches of licensing and certificating. It is believed that teachers' performance and beliefs can have significant effects on the performance and achievement of the learners.

It is demonstrated that the students who pass the course with low-effective teachers receive lower achievement compared with those who are assigned by effective teachers (Sanders & Rivers, 1996). Regarding these points, enhancing students' achievement is highly interwoven with teacher performance in the class. Thus, evaluating and assessing teachers should be an inseparable part of any educational system.

Different variables of teachers' competency are assumed to have a direct relationship with students' achievement and learning. These variables have been the topic of many studies during the previous decades (Darling-Hammond, 2000). Some of them can be named as: subject Matter Knowledge, teaching and learning knowledge, teaching experience, and certification statues.

Teachers are the significant statues that improve the efficacy of students. Therefore, it is the responsibility of principles and supervisors to choose the "highly skilled and motivated teachers" (OECD, 2009). To reach this goal, having continuous monitoring and evaluation should be assigned. Raising educational standards is interconnected with knowing the teachers' strengths and practicing on their weaknesses for having further development and progress (OECD, 2009).

1.2 Statement of the Problem

The common tools for measuring the effect of teachers on the learners' achievement are statistical models. According to Hall, Diaz-Bilello, & Marion (2015), the most highly-profiled models are Value-Added Measures and Student Growth Percentiles. The rationale behind these two models is determining the teachers' effectiveness based on students' standardized assessment (Marthus, 2017). In 2010, Baker declared that applying the statistical models is one of the multiple sources of measurement for evaluating teachers. This has resulted in the lack of enough research on other criteria. Therefore, the current study will focus on the psychometric features in order to evaluate the teacher performance in the class. In Iran, the most widely-used method for evaluating the effectiveness of teachers is applying the statistical measurements. Since learning is a complex concept that is tightly interwoven with various factors, sticking to students' scores would not be a scientific method. As a result, the writer believes that English teachers in institutes should be evaluated by psychometric methods as well. One of the techniques that have been selected for this research is classroom observation. In English institutes, observations are unbelievably skin-deep. In other words, unsystematic and unplanned evaluations are run. The evaluator observes the class once or twice in a term. Unfortunately the English evaluators do not check the teacher performance after providing remedies and feedback in order to find any practicality of their evaluations. Moreover, the students' reactions are not welcomed by any supervisor. Numerous studies have revealed that the process of teacher evaluation would be ineffective when students who are the crucial factors of the educational system, are ignored and passive (The New Teacher Project, 2010).

As a result, in this study observing regularly, giving appropriate feedback and checking the necessary changes are the major concerns. In addition, teacher portfolio which is the best representative of a teacher reflection has been utilized in this study.

1.3 Significance of the Study

The main participants of this study were teachers. The common procedure in order to be accepted as an English teacher in Iran is passing the oral interview, taking written tests which are administrated by the institutes' supervisors, participating in Teacher Training Course (TTC), and after being accepted in the demonstration session, the applicant will be qualified as an English teacher. Thus, classes will be given to them according to the policies of the institutes.

The available evidence seems to suggest that the classes are observed during a term. Different methods are used such as: online observations (sessions will be recorded and analyzed by the principal or observer or head of the department), observer evaluates the class directly (the time of a direct observation is vary according to different institutes and policies), and asking the English learners through questionnaire (this method is not that much common).

Although all these methods are in use, being score-based is a decisive factor in determining the effectiveness of a teacher and a student progress. Regarding this point, Margaret (2017) criticized the idea of "test-based accountability"; therefore, this study makes a major contribution to the field of teacher evaluation by utilizing classroom observation and teacher-portfolio evaluation to find the effect of teacher evaluation on student achievement.

Additionally, this study stresses the point that evaluating a teacher is an on-going process that needs conspicuous consideration.

1.4 Research Questions and Hypotheses

The following questions are the major issues of this study to be determined:

1. Does teacher portfolio-evaluation have any significant effect on Iranian intermediate EFL learners' general language proficiency achievement?
2. Does classroom observation as an evaluation technique have any significant effect on Iranian intermediate EFL learners' general language proficiency achievement?
3. Do teacher portfolio-evaluation and classroom observation have the same significant effect on Iranian intermediate EFL learners' general language proficiency achievement?

2. Review of the Literature

2.1 The Historical Background for Teacher Evaluation

A large and growing body of literature has investigated that "teaching has existed long before teacher evaluation" (Labaree, 2008, p.291). Traditionally, teacher preparation was not required but familiarity with the subject matter was the necessary factor (Labaree, 2008). Therefore, monitoring and evaluating teachers and their teaching techniques were not welcomed. In recent years, there has been an increasing amount of studies on professionalism of educational fields (Marzano et al., 2011; Labree, 2008). Regarding this point Marzano et al. (2011) proposed that after WWII era there has been a critical shift from the industrialized view of education to the focus on teachers and teaching quality which is known as *clinical supervision*. This method was defined as a

"close, helping relationship" between the teacher and supervisors (Okafor, 2012, p.1). This close relationship was examined with the purpose of improving the classroom achievement (Goldhammer, Anderson & Krajewski, 1980). Meanwhile, several studies became interested in the relationship which exists between the behavior of teachers and the students' achievement (Danielson & McGreal, 2000, p.14). Since then, the focus of abundant experts have devoted to the impact of the teachers on students (Rivkn, Hanushek, & Kain, 2005; Rockoff, Jocob, Kane, & Staiger, 2008; Wright, Horn, & Sanders, 1997). One of the prominent models is presented by Danielson. Danielson in 1996 developed a *Framework for Teaching* (FFT model) in which he presented a "clear and meaningful conversation about effective teaching practice" (Danielson Group, 2013, para.1). The focal point which was recited by *Framework for Teaching* was related to the quality of evaluation systems; furthermore, the definition of an acceptable teaching was investigated. Danielson in his book proposed that "everyone in this system such as teachers, mentors, coaches, and supervisors must possess a shared understanding" (p. 35).

2.2 The New Shift in Teacher Evaluation

The scope of evaluation changed during those periods and it seemed that not only the supervision of teachers and their behaviors were the main focus but also the quality and evaluation of teaching and their relation to students' achievement were the centre of attention (Marzano et al., 2011). Before this time, the procedure of evaluation was not scientific and frequent. The supervisor or the observer evaluated the class based on the fixed checklist in order to monitor some specific observable behaviors and features. Moreover, this process was not taken seriously by the teachers and administrators (Danielson & McGreal, 2000; Marzano et al., 2011: Ravitch, 2010; The New Teacher Project, 2010).Several researchers declared that although teacher evaluation is valid and required activity for evaluating the performance of a teacher, classroom-observation-based evaluations can be considered "at best incomplete measures of teaching that produce gains in student achievement and attainment" (Taylor & Tyler, 2011, p.7).Numerous criticisms have been suggested about observation-based evaluation; however, experts and pioneers asserted that a well-designed evaluation system which is connected with planned and frequent observations can have a decisive impact on the development and improvement of students' academic achievement (Taylor & Tylor, 2012, The New Teacher Project, 2010). The same idea recommended by Grissom and Youngs (2016). They believed that "rigorous teacher evaluation systems have the potential to promote the student achievement, but only if the systems are carefully designed and implemented; moreover, the data they generate are interpreted and used properly" (p.2).

2.3 Evaluating Teacher Effectiveness and the Methods

Many evaluation systems have been implemented by several administrators such as "teacher observation rating, student survey feedback, and statistical estimation of a teacher's value-added impact on student achievement" (Grissom and Youngs, 2016; p.1). Goe, Holdheide, L, 2011, represented the instruments that have the potentiality to improve the academic achievements. These tools can be: observation instruments, performance rubrics, portfolios, teacher self-assessment, and parent/student surveys (p.20). Various methods which are prominent in the field of teacher evaluation will be discussed in this study:

Classroom Observation: Reviewing the various studies is tightly interwoven with the traditional and the most well known method which is observation for evaluating teachers (Berk, 2005; Goe & Croft, 2009; Little et al., 2009; Mather, Olivia, & Laine, 2008). The report which has been submitted by Steinberg and Donaldson (2014) suggested that although focusing on the students' scores for evaluating teachers is widely-used, many teachers are evaluated through observation sessions. In addition, it is believed that credibility is a point which is received by many teachers through observation sessions (Little et al., 2009). Conversely, Weisberg et al. (2009) asserted that observation is not an appropriate method for making difference among teachers. The same idea has been proposed by Goe and Croft (2009) that "classroom observations provide a useful measure of teachers' practice but little evidence about whether students are actually learning" (p.5). Moreover, Cohan & Goldhaber (2016) argued that "classroom observations have strong face validity because they assess' process or teaching variables, not student outcomes, which may feel distal from teachers' work" (p.9).

Teacher Portfolios: Another complementary section for evaluating teachers is teacher portfolio. This method is worth doing especially when availability of evaluating the student growth through standardized test score is impossible (Mathus, 2016). According to Little et al. (2009) teacher portfolios are "a collection of materials for the purpose of providing evidence of teachers' practice and student achievement". The materials that can be included in the portfolio are: lesson plans, assessments, student work samples, professional learning or course work, and personal reflections (Berk, 2005; Little et al., 2009). Portfolios should be collected comprehensively and appropriately in a way that provide sufficient information about the "ongoing process and processes that contributed to one's student achievement" (Mathus, 2017).

The Use of Multiple Measurements: Through different studies and researches, several experts and scholars asserted that teaching and learning are complex phenomenon; therefore, teacher assessment is not a unilateral action and it needs multiple measurement tools (Darlind-Hammond et al., 2013; Goe et al., 2008). Regarding this

point, a great bulk of study indicates the fact that applying different methods for evaluating teachers will result in strong and valid decision making for future (Darling-Hammond, 2013; Goe & Holdheide, 2011; Hansen, Lemke & Sorensen, 2013; Henry & Guthrie, 2016; Kane & Staiger, 2012). Moreover, it is encouraged by many researchers that the use of a "combination of formative and summative measures to inform both short and long term professional growth plans" (Burnett et al., 2012, p.5).

2.4 Definition of Language Proficiency

Dealing with the definition of language proficiency has been a major issue in all around the world among language experts. Some experts believe that the language proficiency of an individual can be determined through the person's competency in a foreign language not in the classroom. In other words, the person should be able to read and interpret signs; moreover, interact with the native speakers without any interruptions. Additionally, some of the scholars asserted that language proficiency is the ability to speak and perform in a language. The predominant framework in United States declared that proficient speakers should have the ability to use the language accurately and fluently; in addition, be able to use the variety of discourse strategy

This research was designed to stress the point that language researchers openly confess this ambiguity has not been decoded yet. As what Cummins (1984) declared that the true nature of language proficiency is defined by some researchers in a way that it consists of 64 separate language components while others believed that it consists of only one global factor. The common point in all conceptions or definitions of language proficiency is firstly four basic skills: speaking, reading. Listening, writing; secondly every definition considers language proficiency within a specific context. Therefore, an acceptable English language proficiency test should establish a context that is as nearly as possible to the target language (Del Vacchio and Guerrero, 1995).

3 Methodology

3.1 Participants

The participants of this study were two groups. Because of the nature of this research the main participants were the teachers who had more than ten years of experience in teaching English to Iranian English learners. Another group of participants included the learners who were all at intermediate level and had studied English for about 2-3 years. The institute administrated Oxford Placement Test in order to determine the students' levels. Out of 90 participants 60 students were selected as being at intermediate level. Therefore, in order to exclude upper-intermediate and lower-intermediate candidates initially 60 participants took PET as a homogeneity test. The results left 43 homogeneous participants at the intermediate level which were randomly assigned into three groups (portfolios = 14; observation = 15; control = 14).Also, thirty learners with almost same characteristics of the participants in the main study participated in a pilot study of PET in order to determine the internal consistency of the test as well as inter-rater reliability of the writing and speaking sections. Focusing on the learners' background was beyond the scope of this study. These participants were divided into three groups. Group one was the control group which consisted of 14 learners. The second group which was an experimental group and its teacher was evaluated through observation method included 15 learners. Finally, the last group that consisted of 14 learners was the second experimental group whose teacher was given teacher portfolio treatment.

3.2 Treatment

Treatment sessions were employed for the two experimental groups in two semesters but the point which should be taken into the consideration is the change of treatment for each experimental group in order to give the opportunity of receiving both of the utilized methods for each teacher. As a matter of fact, one of the teachers dealt with classroom observation and the other worked on portfolios in the first semester; in a similar fashion, for the second term the use of methods changed, meaning that the first teacher dealt with portfolio and the second teacher was evaluated through classroom observation. The aim of this substitution was creating a parallel situation for both experimental groups additionally, enhancing the validity of the research result. This could help to minimize any irrelevant teacher effect.

3.3 Instruments and Materials

The following instruments were employed in order to collect the data:

Oxford Placement Test: The second version of Oxford Placement Test (OPT) was administrated with the aim of selecting intermediate students. Oxford Placement Test consists of multiple choice items, having 60 questions in grammar, reading, and vocabulary with one point for each correct item. The allocated time was 40 minutes

PET: In the present study thorough Oxford Placement Test, intermediate learners were diagnosed but intermediate level itself is divided in lower intermediate, intermediate, and upper intermediate; therefore, PET was administrated in order to demonstrate the true intermediate participants. The adult version was used for this study.

IELTS General Proficiency Test: Because the ultimate goal of this study was to determine the general

language proficiency achievement of the learners, the most effective measurement for this purpose could be IELTS general proficiency tests. IELTS stands for "International English Language Testing System." The prime aim of this study is to assess the English proficiency of non-native speakers. Having the general language proficiency is the ability to speak, read, listen, and write accurately and comprehensively. Therefore, IELTS general tests can be one of the best measurement instruments that can check the general ability of the learners in all four main skills and sub-skills. The latest IELTS general training test (the twelfth edition) was used for this study.

Portfolio: One of the techniques that were incorporated in this study was teacher portfolio. Portfolios are unique by their nature due to the way they are collected and created by the teachers for the purpose of evaluation and example of their practice. The ultimate purpose of teacher portfolio is to exhibit the classroom procedure and to fulfill the predetermined standards that have been established by the supervisors. In this study teachers were expected to include teaching goal, syllabus, students' home works, teacher assessment, teachers' comment on students' progress, and the utilized materials during the course. The evaluator analyzed the lesson each session and provided the teachers with appropriate and detailed remedies and feedbacks.

Classroom observation: In this study the evaluator observed the classes for about 30 minutes each session. The prime purpose was to analyze and evaluate the performance of the teachers based on the criteria which are proposed by Danielson's *Framework for Teaching*. According to this book, the observation process can be beneficial if three phases are met. These three phases are pre-observation phase which includes interviewing with teachers about their strategies, sharing any specific characteristics of learners and learning environment, classroom observation, and post-observation phase which consists of cooperation between the teacher and the observer in order to remove the weaknesses and promote the strengths**.**

Teaching Knowledge Test: Teaching knowledge test was administrated with the purpose of determining the teaching knowledge of the teachers. It consists of 240 questions (80 questions for each module). This test focuses on the concept of teaching but solely on the abstract concepts. In other words, the practical part of teaching has been removed. Teaching knowledge test consists of multiple-choice tests and it has three modules that can be taken separately or together.

4. Results

4.1 Data Analysis

Table 1 displays the result of PET in order to select the participants of the study; the researcher selected those individuals whose PET scores fell within the range of -1 SD and +1 SD (65.31 to 84.73). Following this procedure resulted in keeping 43 individuals as the homogenous participants of the study. Table 1 presents the descriptive statistics pertinent to the remaining 43 test takers.

Table 1:
Descriptive Statistics for the Homogenous Participants

	N	Minimum	Maximum	Mean	SD	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Selected	43	66.50	83.50	75.9767	4.45874	-.105	.361	-.822	.709
Valid N (listwise)	43								

The 43 participants who were selected as the homogeneous participants of the study regarding their language proficiency were then assigned into two experimental and a control groups.

Homogeneity in terms of teaching knowledge: three teachers instructed the participants in two semesters. In order to make sure that these three teachers (two teachers were for experimental groups and one teacher was selected for control group) were homogenous with regards to teaching knowledge, they were asked to sit in a Teaching Knowledge Test (TKT) before the treatment initiated. Table 2 shows the results obtained by teachers in each module of the test.

Table 2:
Scores Obtained from TKT

		Module		
		1.00	2.00	3.00
TKT	First Teacher	54.00	64.00	66.00
	Second Teacher	58.00	63.00	67.00
	Third Teacher	56.00	62.00	65.00

A one-way ANOVA was run on these scores to see if there is any significant difference between the knowledge of the two teachers (Table 3).

Table 3:
ANOVA: the Difference between Two Teachers' TKT

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	2.667	1	2.667	.086	.783
Within Groups	123.333	4	30.833		
Total	126.000	5			

As the results show, the two teachers had a very low difference (F=.08. p = .78) with regards to their teaching knowledge.

An IELTS test was administered to the participants in order to see if they are in the same level of knowledge. The listening and reading sections of the test were mostly multiple-choice or short answers with an answer key, each including 40 items. However, as both writing and speaking of the test had to be rated based on a rating scale, about 25 percent of the papers (10 cases) were rated by two raters and their inter-rater consistency were assured before continuing the measurement. Table 4 shows the results.

Table 4:
Kappa: Inter-Rater Agreement for the Writing and Speaking Sections of IELTS

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Writing	Kappa	.538	.202	2.577	.010
Speaking	Kappa	.623	.212	2.414	.016
N of Valid Cases		10			

The strong indices of Kappa agreement coefficients (.54 and .62) for both writing and speaking were indicator of high consistency between the two raters' scorings. Therefore, the researcher was rest assured that the two raters can continue rating the participants' performances. After making sure of the rating consistency, the participants' performances were scored. In order to simplify the analysis, the band scores of the participants were used in the data analysis. Table 5 presents the three groups' scores in IELTS pretest.

Table 5:
Descriptive Statistics of Pretest Scores by Three Groups

	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
Ex1	14	4.0000	.65044	.17384	3.00	5.50
Ex2	15	4.0667	.62297	.16085	3.00	5.00
Control	14	4.1071	.59416	.15880	3.00	5.50
Total	43	4.0581	.60954	.09295	3.00	5.50

In order to check if there is any initial significant difference among the three groups, a one-way ANOVA was run. Before running ANOVA, the homogeneity of error variances was checked (Table 6).

Table 6:
Levene's Test of Homogeneity of Variances: Pretest Scores

Levene Statistic	df1	df2	Sig.
.082	2	40	.921

As the results of levene's test ($F_{(2,40)} = .08$, $p = .921 > .05$) confirmed, the variances among the three groups were not significantly different; thus the assumption of homogeneity of variances was met. Table 7 presents the results of ANOVA.

Table7:
ANOVA: Pretest Scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.082	2	.041	.106	.900
Within Groups	15.523	40	.388		
Total	15.605	42			

Table 8 shows the descriptive statistics of the posttest scores.

Table 8:
Descriptive Statistics of Posttest Scores by Three Groups

	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
Ex1	14	4.6071	.44629	.11928	4.00	5.50
Ex2	15	4.6667	.40825	.10541	4.00	5.00
Control	14	4.1429	.60219	.16094	3.00	5.50
Total	43	4.4767	.53400	.08143	3.00	5.50

As it is evident in Table 8, while the two experimental groups have almost the same mean scores, the mean score of the control group is different. In order to see if this difference is significant, a one-way ANOVA was run. Before running ANOVA, the assumption of homogeneity of variances was checked (Table 9).

Table 9:
Levene's Test of Homogeneity of Variances: Posttest Scores

Levene Statistic	df1	df2	Sig.
.779	2	40	.466

The results of Levene's test (F(2,40) = .779, p = .47 >.05) shows that the assumption is met. Table 10 presents the result of ANOVA.

Table 10:
ANOVA: Posttest Scores

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	2.340	2	1.170	4.856	.013
Within Groups	9.637	40	.241		
Total	11.977	42			

The results of ANOVA ($F_{(2,42)}$ = 4.85, p = .013 < .05) indicates that the difference between the means is significant.. In order to locate the place of difference, a Tukey post hoc was run (Table 11).

Table 11:
Multiple Comparisons: Tukey Post Hoc on Posttest Scores

		Mean Difference			95% Confidence Interval	
(I) Group	(J) Group	(I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Ex1	Ex2	-.05952	.18240	.943	-.5035	.3844
	Control	.46429*	.18552	.043	.0127	.9158
Ex2	Ex1	.05952	.18240	.943	-.3844	.5035
	Control	.52381*	.18240	.017	.0799	.9678
Control	Ex1	-.46429*	.18552	.043	-.9158	-.0127
	Ex2	-.52381*	.18240	.017	-.9678	-.0799

Therefore, as the results of the tables indicated teacher-portfolio evaluation and classroom observation had a significant effect on the general language proficiency achievement of the two experimental groups.

5. Discussion

Turning back to the reviewing of the literature, the researcher asserted that the study investigated the effect of teacher evaluation with the combination of two instruments: classroom observation and teacher portfolio on language proficiency achievement of the EFL learners have not been carried out. This study has tried to bridge this gap in this field of knowledge. Considering all these issues and discussions, we turn back to the questions of this study:

1. Does teacher portfolio-evaluation as have any significant effect on Iranian intermediate EFL learners' general language proficiency achievement?

According to the information which has been displayed in table 11 the first experimental group that received teacher portfolio had significantly higher mean score (MD = .46, SE = .19, p = .04 < .05) than control group, indicating that the treatment was effective. The researchers did not find any pertinent studies that were truly in line with the current study that investigated the significant effect of teacher-portfolio evaluation on the general language proficiency achievement of the learners. The innovativeness of this study with this specific and useful method with the aim of improving students' general language proficiency achievement can motivate numerous scholars for more investigations.

1. Does classroom observation as an evaluation technique have any significant effect on Iranian intermediate EFL learners' general language proficiency achievement?

Additionally, the second experimental in which classroom observation was carried out had noticeable higher

mean as well (MD = .52, SE = .18, p = .02 < .05) than control group, it indicated that this treatment was also effective. Goe and Croft (2009) in their study strongly asserted that "Classroom observations provide a useful measure of teachers' practice but little evidence about whether students are actually learning" (p. 5). However, in this study the researcher proved that there is a direct relationship between classroom observation and learners' actual learning.

2. Do teacher portfolio-evaluation and classroom observation have the same significant effect on Iranian intermediate EFL learners' general language proficiency achievement?

Regarding the third question which focused on the same significant effect of teacher portfolio-evaluation and classroom observation, the table 11 shows that the post-test scores of the two experimental groups was very close to each other (MD = .06, SE = .18, p = .94 > .05). Therefore, this result revealed that the two treatments did not have different effects upon the language proficiency achievement of the learners and their effectiveness was almost the same.

6. Conclusion, pedagogical implications, and future studies

It is worth mentioning that this study is the first research that explores the effect of teacher evaluation on the learners' general language proficiency achievement by the use of classroom observation and teacher portfolio. Learning and teaching are complex phenomena; thus, researching in these areas needs several investigations and considerations. In other words, isolation of teaching from learning or learning from teaching is almost impossible. This point can be worse if the researcher deals with the achievement of the learners. As it has been mentioned by Bracey (2006) "Peer interactions, school climate, and the nonrandom placement of students all contribute to student achievement in significant and unquantifiable ways".

In this study the researcher tried to investigate the role of teacher evaluation process in English institutes. The final purpose was to maximize the English learners' outcome. The dominant criteria at English institutes for enhancing the academic achievement of the learners are focusing on the teachers' performance through classroom observations which is apparently non-academic and unsystematic. As a matter of fact, the authorities, observers, and teachers and more importantly the learners cannot meet the usefulness of this technique in the process of teaching and learning. Moreover, it can be concluded that this evaluation system cannot have any significant effect on the academic achievement of the learners.

The current study selected these two techniques among many measuring instruments that are in use in several teacher evaluation systems but the point which should be taken into the consideration is that "unfortunately, there is little empirical evidence of the validity of these various methods for measuring the effectiveness of teachers in teacher evaluation process, and in many cases, there are no standardized instruments for data collection. Instead, the collection of data-and decisions about what is important to collect-is left up to authorities" (Goe, Bell, & Little, 2008; p.5). In this case the writers combined the two widely-used techniques to maintain the validity and practicality of the result. Since the main participant of this study were teachers, the role of the learners and the factors that are related to their achievement were beyond the scope of this study; therefore, they were not investigated fully. The results of the study revealed that both of the techniques had a significant effect on the general language proficiency achievement of the learners. However the point which is worth mentioning is that further studies and researchers should shed light on the effect of different teacher evaluation methods and their combinations in order to improve the outcome of the learners in different contexts.

Various studies have investigated the effect of teacher evaluation on the achievement of the learners and the techniques they applied were mostly the use of one measuring instrument. This study is tremendously new in the field of teacher evaluation that incorporated two techniques together. The findings of this study present vital pedagogical implications for any educational system and more importantly for policy makers' future decisions. As a matter of fact, this study is broadly in line with those of researchers such as Danielson (2001), Hattie (2012), Grant, Hindman, & Stronge (2013), Mathus (2017) who asserted that the factor which has the greatest influence on the achievement and learning of the students is teacher performance and effectiveness. Therefore, it was concluded that if authorities put more stress on the role of teachers and their strategies consistently and systematically, Iranian English learners can meet the most academic achievement in final run. One issue which is worth knowing is that Iranian English learners are living in a non-English environment. In other words, English is considered as a second language for them; therefore, enriching the evaluation system should get priority to enhance the achievement of the English learners.

Several researchers in different parts of the world investigated the relationship or effect of teacher evaluation with the learners' achievement in school-contexts. The present study tried to examine this issue in English institutes in Iran which is extremely new and fresh. The same method can be transferred to Iran's public schools and even for other subjects.

References

Berk, R. A., (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching*

and Learning in Higher Education. 17(1), 48- 62. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1. 454.3400&rep=rep1&type. Pdf

Bracey, G. W. (2006).Value-Added Models, Front and Center. *Phi Delta Kappan, 87*(6), 478.

Burnett, A., Cushing, E., & Bivona, L. (2012). Uses of multiple measures for performance-based comprehension. *Center for Educator Compensation Reform.* Retrieved from http://files.eric.ed.gov/fulltext/ED533704.pdf

Cohen, J., & Goldhaber, D. (2016). Observations on evaluating teacher performance: Assessing the strengths and weaknesses of classroom observations and value added measures. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems* (pp. 8-21). New York, NY: Teachers College, Columbia University.

Cummins, J. (1984). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles: National Dissemination and Assessment Center.

Danielson, C. & McGreal, T. L. (2000). *Teacher evaluation: To enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson Group. (2013).*The framework.* Retrieved from the Danielson Group website: http://danielsongroup.org/framework/

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). Retrieved from http://epaa.asu.edu/ojs/article/view/392/515.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement?* New Your, NY: Teachers College Press.

Del Vecchio, A. & Guerrero, M. (1995). *Handbook of English language proficiency tests.* Albuquerque, NM: New Mexico Highlands University, Evaluation Assistance Center–Western Region.

Ellet, C. D., & Teddle, C., 2003. *Teacher Evaluation, Teacher Effectiveness and School Effectiveness: Perspectives from the USA*. Journal of Personnel Evaluation in Education, 17(1), p. 101-128, Kluwer Academic Publishers. Manufactured in the Netherlands

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness. Research-to-Practice Brief*. National Comfrehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/RestoPractice_Evaluating TeacherEffectiveness.pdf

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects* [Research & Policy Brief]. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved fromhttp://www.gtlcenter.org/sites/default/files/docs/MeasuringTeachersContributions.pdf

Goldhammer, R., Anderson, R. H. & Krajewski, R. J. (1980). *Clinical supervision: Special methods for the supervision of teachers,* 2nd ed. New York, NY: Holt, Rinehart, and Winston, Inc.

Grissom, J. A., & Youngs, P. (2016). Making the most of multiple measures. In J. A.Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems* (pp. 1-7). New York, NY: Teachers College Press.

Hall, E., Diaz-Bilello, E., & Marion, S. (2015). *Considerations for establishing performance standards for educator evaluation systems*. Retrieved from http:// www.nciea.org/publication_PDFs/Establishing%20Performance%20Standards%2 0for%20EES_EH2015.pdf

Hansen, M., Lemke, M., & Sorensen, N. (2013). *Combining multiple performance measures: Do common approaches undermine districts' personnel evaluation systems?* Retrieved from American Institutes for Research website:http://www.air.org

Hattie, J., & Anderman, E. M. (Eds). (2013). *International guide to student achievement.* New York, NY: Routledge.

Henry, G. T., & Guthrie, J. E. (2016). Using multiple measures for development teacher evaluation. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems* (pp. 143-155). New York, NY: Teachers College, Columbia University.

Kane, T. J., & Staiger, D. O. (2012, January). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* [Research paper]. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Labaree, D. F. (2008). *An uneasy relationship: the history of teach eduation in the university*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.462.6342&rep=rep1&type=pdf

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. *National Comprehensive Center for Teacher Quality*. Retrieved from http://files.eric.ed.gov/fulltext/ED543776.pdf.

Marzano, R. J., Frontier, T., & Livingstone, D. (2011). *Effective supervision: Supporting the art and science of teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Mathers, C., Olivia, M., & Laine, S. W. M. (2008). *Improving instruction through effective teacher evaluation: Options for states and districts* [TQ Research & Policy Brief]. Washington, DC: National Comprehensive

Center for Teacher Quality. Retrieved from http://files.eric.ed.gov/fulltext/ED520778.pdf

Mathus, M., 2017. The Relationship between Teacher Evaluation Ratings and Student Achievement in a Rural, Midwest School District. A Dissertation submitted to the Education Faculty of Lindenwood University.

OECD (2009a), *OECD Review on Evaluation and Assessment for Improving School Outcomes: Design and Implementation Plan for the Review*, OECD, Paris [OLIS Document EDU/ EDPC (2009)3/REV1].

Okafor, P. (2012). *Leadersip in instructional supervision*. Retrieved from http://patrick okafor.com/files/ClinicalSupervision.pdf

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*(2), 417-458. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.322.4872&rep=rep1&type=pdf

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one?* (NBER Working Paper No. 14485).

Sanders, W.L. and Rivers, J.C. (1996). Cumulative and residual effects of teachers on future student academic achievement. Knoxville, TN. University of Tennessee Value-Added Research and Assessment Center.

Steinberg, M. & Donaldson, M. (2014). *The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era*. [Policy Brief]. Retrieved from http://cepa.uconn.edu/wp-content/uploads/sites/

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers*. (NBER Working Paper No. 16877).National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16877.pdf

Taylor, E. S. & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-51. Doi: 10.1257/aer.102.7.3628

The New Teacher Project. (2010, May 25). *Teacher Evaluation 2.0.* Boulder, CO: National Education Policy Center. Retrieved from http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Brooklyn, NY: New Teacher Project. Retrieved from http://tntp.org/publications/view/the-widget-effect-failure-to-act-on-differences-in-teacher-effectiveness