Issue in Language Teaching (ILT), Vol. 7, No. 2, 147-187, December 2018

Teacher Evaluation in EFL Context: Development and Validation of a Teacher Evaluation Questionnaire

Masoomeh Estaji*
Associate Professor of Applied Linguistics, Allameh Tabataba'i University
Mohsen Shafaghi
PhD Candidate of TEFL, Allameh Tabataba'i University

Abstract

Teacher Evaluation (TE) is a critical and controversial process in the teaching profession and formal education system. Effective TE requires both sound policy implementation and efficient processes, affecting the efficiency of the education system. To present a framework for research and highlight the constructs of TE, this study developed and validated a teacher evaluation questionnaire. To this end, seven TE components were identified after undertaking a comprehensive review of the literature and conducting interviews with domain experts on TE. Then a draft version of the TE questionnaire, consisting of 105 items, was pilot tested with 330 teacher evaluators, who were working for various English language institutes in Iran. The results, using Exploratory Factor Analysis (EFA), led to a 90-item questionnaire with strong estimates of reliability and validity. The results also demonstrated that the questionnaire consisted of a six-factor structure of perception, method, system, content, purpose, and outcome of TE. The subsequent Confirmatory Factor Analysis (CFA) of the data from another 360 Iranian teacher evaluators, selected through convenience sampling, indicated that the six-factor structure of the questionnaire was statistically supported, meaning that the questionnaire's detected constructs were not the result of random variance in the participants' responses. The results of the study have presented a framework for research and highlighted the principles of teacher evaluation.

Keywords: Confirmatory Factor Analysis, Exploratory Factor Analysis, evaluator's perceptions, teacher evaluation, teacher evaluation systems

***Corresponding author's email:** mestaji74@gmail.com

INTRODUCTION

Evaluation is directly associated with the teaching profession and is at the heart of teaching performance which has given rise to the term "*Teacher Evaluation*". Teacher Evaluation (TE) has been one of the central concerns in teacher education programs. Many countries around the world have implemented teacher evaluation programs to improve their teaching systems (Delvaux et al., 2013). However, there is diversity in teacher evaluation systems worldwide. Teacher evaluation and development policies around the world are undergoing a significant change. Despite the lack of consensus, any country's education system requires some standards to meet. Prasertsin (2015) conducted a study on teaching standards and through confirmatory factor analysis found research, evaluation, measurement, and quality assurance as fundamental factors on education improvement.

In English language institutes, teachers normally work together in the same environment. During their professional lives, they may be subject to various teacher appraisal judgments by a classroom observer from the same institute or an unfamiliar observer from a different center. There might be cases in which supervisors do not observe the type of lesson that the teacher has taught and this might lead to an invalid judgment. Thus, observers need to display objectivity, bearing in mind the existing differences in training and background (Howard, 2010).

The necessity for teacher evaluation in teacher education programs can be explored from diverse perspectives. Firstly, observation and evaluation are critical endeavors which can cause life-long consequences in one's life. Likewise, not everyone has enough time or expertise to evaluate how teachers actually work. Furthermore, the tools to document and evaluate teachers' perceptions and performance are deficient. Even if there are tools and instruments to evaluate teachers, the implementation, interpretations, and inferences derived from them are typically invalid and inadequate. More importantly, despite the various observation tools and systems available for teacher evaluation, the research literature lacks a strong and attested theory and method to evaluate teachers' performance. Thus, there is an ongoing debate on how best to evaluate teacher performance.

Teacher education programs propose that evaluation must form a culture of continuous learning for teachers and evaluators. Continuous learning and practicing is a way toward Professional Development (PD). One of the components of PD is using trained individuals to evaluate and provide feedback (Goe, Biggers, & Croft, 2012). Therefore, evaluation can be viewed as one of the components of PD for teachers. However, there are occasions where teachers are not improving or are unable to improve. Teachers may be effective in some areas, but may still have a long way to growth. Teacher evaluators can have a beneficial role in this way if they truly judge the teachers' performance and help them perform more effectively.

The evaluators and decision-makers should be capable of providing constructive feedback to teachers. The most remarkable teacher evaluation system can be of little value if the evaluator and decision-makers are not supportive. Regarding teacher evaluation systems, Kyriakides, Demetriou, and Charalambous (2006) argue that the development of a valid TE system is problematic in many educational systems. This problem might stem from a lack of empirical evidence regarding the characteristics of effective teacher evaluation systems. Further, there are still schools and private language centers which are traditionally hierarchical in structure and management. Consequently, teachers need to work closely with their respective evaluators in resolving problems resulting from their classroom instruction and student management in their daily work practice. To do so, they attempt to please their supervisors as they are worried about the consequences of being rated as "unsatisfactory" or even being fired (Moradi, Sepehrifar, & Parhizkar, 2014).

Accordingly, teacher evaluation is extremely significant and should be regarded as a systematic process. The findings of this study might be interesting to anyone seeking for the status quo of teacher evaluation in Iran. In addition, it can contribute to pre-service student teachers in teacher training centers, in-service teachers, teacher educators, teacher evaluators, and teacher education policy makers. For this to happen, a standard teacher evaluation instrument appropriate for an EFL context needs to be designed. Besides, the reliability and validity of the evaluation instrument should be confirmed. To address the above-mentioned issues, the present study aimed to examine the perceptions of evaluators, the criteria, the methods, and systems they have employed in the evaluation of English language teachers' performance, and to develop and validate a teacher evaluation questionnaire.

LITERATURE REVIEW

The literature review in this study goes beyond the EFL context to research studies in teacher education. Historically, the background of teacher evaluation can be traced back from the turn of the twentieth century to about 1980s. This history can be divided into three coinciding periods: (a) looking for great teachers; (b) determining teacher quality based on student learning; and (c) analyzing pedagogical practices. At the beginning of the twenty-first century, there has been a transition in teacher evaluation to a period of Evaluating Teaching as a Professional Behavior (Medley, Coker, & Soar, 1984).

Traditionally, teacher evaluation and professional development have been informed by two different models: The standards-based model which underscores the use of explicit frameworks to model classroom practice and quality instruction (Peterson, 2000), and the outcome-based model which privileges productivity in terms of student achievement and other relevant outcomes (Kennedy, 2010). Recently, these two models have commenced to converge, as a new product of teacher evaluation systems and development combines, which emphasizes the learner gain with explicit and detailed models of teaching practice (Kane & Staiger, 2012). In these two models, two lines of thought have emerged; the first one supports evaluation as an essential component of the profession; however, it has had limited impact on teacher growth and student performance. The second places an emphasis on the difference between summative and formative evaluations.

Concerning evaluation systems, it is crucial to raise evaluators' awareness of various evaluation models and systems. Numerous studies have attempted to design models of teacher evaluation. For instance, Bryant, Maarouf, Burcham, and Greer (2016) working on Danielson's (2013) framework for teacher assessment, aimed to confirm the quality of the framework through examining the internal consistency and its construct validity. The study presented a 16-item framework with high reliability and validity in four teaching domains including planning and preparation, the classroom environment, instruction, and professional responsibilities.

Similarly, Ruprich and Urhahne (2015) conducted a similar study by designing a questionnaire for the assessment of teacher goals with 302 teachers in Germany. The results revealed a positive correlation between teachers' goals and observer-assessed classroom management and learning-conducive climate. Further, in a study aiming to explore the Iranian EFL teachers' attitude toward evaluation and its influence on their classroom decision making, Rahmany, Hasani, and Parhoodeh (2014) found that teaching experiences of the teachers obviously affected their attitudes toward teacher evaluation; i.e., less experienced teachers were more influenced by the supervision process whereas more experienced teachers mostly held pessimistic views toward teacher evaluation.

Regarding classroom observation, Wang and Day (2002) found both subjective and procedural problems with observation practices, which affect observer-observee relationship and minimize the role of teachers to submissive performers. As for post-observation sessions, Iyer-O'Sullivan (2015) found that post-observation feedback can be challenging as both the evaluator and teacher would have experienced different emotions and thoughts during observation.

Recent changes in evaluation models have been designed to transform teacher evaluation practices and enhance teacher effectiveness. To this end, Clenchy (2017) conducted a qualitative study to measure the teachers' perceptions of TE models. The results showed that trust, respect, integrity and professionalism were considered as crucial components of effective evaluation models by all participants to foster professional growth. Moreover, it was found that the effectiveness of the evaluator had a significant role in the successful implementation of evaluation models that focused on continuous teacher development. In order to explore EFL teachers' professional challenges, Razmjoo and Mavaddat (2016) developed a model. The results confirmed teacher evaluation as one of the great challenges of the teachers.

The most admirable attempt to identify TE factors has been Martinez, Taut, and Schaaf's (2016) heuristic model which subsumes different variables exploring sixteen classroom observation systems in six countries. Their study presents an analytic framework for orientating with classroom observation and teacher evaluation systems across countries. The framework for systems consists of three fundamental dimensions: 1. conceptual issues (instructional practice and teacher effectiveness); 2. methodological issues (methods used to gather information about these constructs); 3. policy issues (policy, context, processes, and decisions that shape the evaluation).

Martinez et al. (2016) argue that conceptual issues deal with the theoretical or conceptual underpinnings that will provide the basis for understanding, describing, and assessing teacher practice. Moreover, methodological issues are considered in the design and use of observation systems for teacher evaluation and development which include the identity and qualifications of the observers, the modes of observation, and the number and type of observations per teacher. Finally, policy-related factors include the actors involved in creating and designing TE systems, buy-in from key stakeholders, and public narratives around the system's motivation, credibility, and predicted consequences.

PURPOSE OF THE STUDY

Despite the transparency and multidimensionality of the models, mainly Martinez et al.'s (2016) TE analytic framework, they report no reliability and validity index for that. Similar research findings at the international level also verify the ineffectiveness of traditional evaluation models (Kyriakides, Campbell, & Christofidou, 2002). Despite the advances in teacher evaluation models throughout years, evaluation procedures have remained rather unaffected. The need for more reliable and valid measures of TE for pedagogical language contexts thus continues to exist. Considering the instrumental gap in the field, this study attempted to capture the evaluators' perceptions of TE and specify the criteria they have employed in evaluating teachers in order to design and validate a TE questionnaire.

METHOD

Participants

The participants of this study were Iranian EFL teacher evaluators working for private English language institutes. They were of different ages, genders, educational levels, and backgrounds. In this sample, 56% of the participants were female and 44% were male between the ages of 23 and 47 years (M= 35, SD= 1.73). They had a minimum of five years' teaching and evaluation experience. As for their academic degrees, 32% of the participants had Bachelor's, 55% had Master’s, 8% had Doctoral (Ph.D.) degrees, and 5% had a high school diploma or an alternative educational degree. The participants were selected through convenience sampling. The draft version of the TE questionnaire was administered to a group of 330 teacher evaluators in different cities of Iran. The final version of the questionnaire was administered to another group of 360 Iranian teacher evaluators. Overall, a total of 690 teacher evaluators were selected and involved in the completion of the first and final version of the questionnaire.

Instrumentation

In this study, a teacher evaluation questionnaire was designed. To construct the TE questionnaire, items were developed based on the existing questionnaires, review of the related literature, and interviews with experts in the field. The interview questions presented the main constructs of the questionnaire, focusing on the various dimensions of teacher evaluation. In order to avoid any biased item order, the items were randomized in the questionnaire. Further, the purpose of the questionnaire and the way to complete the items were written through clear instructions. The rating scale employed was plainly explained as well. The questionnaire consisted of items on a five-point Likert scale rating from “totally disagree” (rated 1) to “totally agree” (rated 5) as well as one short answer question to give participants choice to write their opinions.

Data Collection Procedure

The process of developing a TE questionnaire in the study followed a standard, step by-step procedure. The questionnaire development started with a careful scrutiny of the related literature on various variables encompassing many perceptional, methodological, and systematic variables. To this end, it was required to develop a pertinent, well-ordered, and flexible framework pertaining to teacher evaluation. Examining the literature on the evaluation of various academic development activities, Chalmers and Gardiner (2015) found the purpose, effectiveness, and impact of teaching preparation programs, the impact of institutional culture, and the measurement approaches as the major themes in the design of an evaluation framework. They suggested that teacher evaluation programs can be divided into three main categories: Teacher focused, student focused, or institutional focused. Among the various teacher evaluation systems available, such as Chalmers and Gardiner's (2015) framework and CIPP evaluation model (Stufflebeam, 1969) focusing on Context, Input, Process, and Product components, the framework presented by Martinez et al. (2016) was applied in this study consisting of conceptual, methodological, and policy issues.

After the framework being chosen, a series of interviews was conducted with 30 supervisor experts (supervising professors, teacher mentors, and supervisors). The purpose of interviews was primarily to see whether the interviewees confirm the variables found significant in TE literature or not. Moreover, they were run to find out whether the interviewees (supervising professors and teacher evaluators) would indicate other important variables relevant to TE. The interviews were semi-structured, beginning with predetermined questions (Appendix A). However, they were not fixed and thus, it was possible for unpredicted questions to emerge during the interview sessions. The data gathered through interviews were content-analyzed based on the guidelines for analyzing the qualitative data (e.g., Mayring, 2014).

Then the results of inductive content analysis of the interviews were examined to approve the variables which were assumed significant for TE. There was a high inter-coder agreement (93%) on the responses and coding of the interview content. Finally, the literature review and content analysis of the interviews led us to the identification of seven components related to teacher evaluation as follows: Perceptions of teacher evaluators, methods of evaluating teachers, teacher evaluation systems, contents of evaluating teachers, purposes of evaluating teachers, outcomes of evaluating teachers, and procedures of teacher evaluation.

Afterward, the existing questionnaires on teacher evaluation were thoroughly analyzed to detect the relevant items. The questionnaires included educational evaluation, teacher evaluation systems and frameworks, supervision and supervisory tasks, summative and formative evaluation, students' evaluation of teacher performance, and appraisal of teachers' performance. Finally, Behlow's (1990) questionnaire, Fyson's (1993) study, and Lowe's (2000) questionnaire were used in this study as guides in order to design our TE questionnaire. As a result, an item pool was developed for all the constructs. Each item was designed based on the related literature, existing questionnaires, and the interviews conducted. Then, based on the focus of the study, the relevant items were selected to be included in the first draft of the questionnaire. Subsequently, the written items were submitted to several supervisors and teacher education experts to judge the redundancy, content validity, clarity of the items, and language. The experts gave their recommendations as to how to modify the potential items which could be misunderstood by the respondents. These steps led to the construction of 105 items that were written by the researchers.

At the end, the developed questionnaire was administered following similar procedures in formal language classrooms in private language institutes. At the time of data collection, the respondents were supervising EFL teachers either in formal language learning classrooms or at the debriefing sessions between the evaluators and the teachers in the private English language institutes. Furthermore, the willingness of the respondents was examined by asking whether they would like to communicate, or be replaced by another respondent.

RESULTS

Reliability Analysis of the Questionnaire

In order to measure the reliability of the questionnaire used in this study, it was administered to 30 teacher evaluators. The responses were submitted to Statistical Packages for Social Sciences (SPSS version 22). Some positive-worded items had been negatively-worded so as to strengthen the reliability of the questionnaire; therefore, the negatively-worded items were reverse-scored before the conduction of further analysis. These pair items were as follows: 2 and 3, 33 and 77, 39 and 40, 66 and 68, 69 and 70, 74 and 75, 81 and 82, 91 and 95, 99 and 103, and 100 and 101. Then Cronbach's alpha was run. Table 1 presents the descriptive summary of the whole questionnaire.

Table 1: The Descriptive Summary of the Whole Questionnaire

Mean	Variance	Std. Deviation	N of Items
391.61	3115.66	55.81	105

As can be seen in Table 1, in totality, the number of items was 105, the mean of the whole questionnaire was 391.61 and the standard deviation was 55.81. Table 2 indicates the overall Cronbach alpha value for this questionnaire.

Table 2: The Results of Cronbach's Alpha for the Questionnaire

Cronbach's Alpha	N of Items
.92	105

The results, as shown in Table 2 above, indicated a satisfactory level of reliability, i.e., α = .92. Moreover, high correlations were found between the responses on each item and the whole questionnaire, except for some items. However, the analysis indicated that omitting these items did not bring about a considerable increase in the reliability of the questionnaire. Therefore, all the items were maintained in the questionnaire.

Validity Analysis of the Questionnaire: Detecting the Factor Structure in TE Questionnaire

Exploratory Factor Analysis (EFA) of the data from the first group of teacher evaluators filling out the draft version of the TE questionnaire was done to determine the factor structure of the TE questionnaire. As a prerequisite for factor analysis, the suitability of data must be investigated. Regarding the sample size required it should be at least 150 cases (Pallant, 2013), which was not violated in this study (N=330). The second assumption is the factorability of the correlation matrix. In order for the data to be regarded as suitable for factor analysis, at least 20% of the correlation matrix and the anti-image correlation matrix must be equal to or greater than .3. Using this matrix, we can identify items that do not correlate with any of the factors. Likewise, the diagonals must be more than .5. After the conduction of the first Principal Component Analysis (PCA) using Promax rotation for the present data, 5 items were excluded based on these two criteria (items 17, 24, 25, 44, and 54). After the omission of these items, another PCA using Promax rotation was run.

Kaiser-Meyer-Olkin's (KMO) measure of sample adequacy and the Bartlett's test of sphericity were also used to check the factorability of the data (Pallant, 2013). The KMO measure of sampling adequacy for the data under study was .92 which was well above the minimum required level of .60 (Tabachnick & Fidell, 2001) and the Bartlett's test of sphericity was significant at p<.001. Both indices supported the factorability of the data. To determine the number of factors, Kaiser's criterion, which claims that eigenvalues must be more than 1, was checked. Table 3 below reports those components whose eigenvalues were above 1.

Table 3: The Results for the Factors with the Eigenvalues of more than 1 in the 2nd PCA

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**
1	24.3	24.31	24.31	24.3	24.31	24.31	\20.73
2	1	12.81	37.13	1	12.81	37.13	14.71
3	12.8	11.35	48.48	12.8	11.35	48.48	13.88
4	1	9.62	58.11	1	9.62	58.11	12.01
5	11.3	8.90	67.01	11.3	8.90	67.01	11.70
6	5	4.66	71.67	5	4.66	71.67	8.41
7	9.62	2.11	73.79	9.62	2.11	73.79	5.32
8	8.90	1.86	75.65	8.90	1.86	75.65	2.97
9	4.66	1.23	76.88	4.66	1.23	76.88	4.07
10	2.11	1.14	78.03	2.11	1.14	78.03	6.42
11	1.86	1.00	79.03	1.86	1.00	79.03	8.20
	1.23			1.23			
	1.14			1.14			
	1.00			1.00			

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

As shown in Table 3, there were 11 components with eigenvalues of more than 1. These components could explain a total of 79.03 percent of the total variance. The first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, and eleventh factors could explain almost 24%, 13%, 11%, 10%, 9%, 5%, 2%, 2%, 1%, 1%, and 1% of the total variance, respectively. As is clear, factor 7 just like factor 8 could only explain 2% of the total variance and factors 9 to 11 could only account for 1% of the total variance, which is deemed to be very low. The decision as to the number of factors to be retained was guided by eigenvalues above 1 and inspection of the scree plot. The scree plot of the second PCA is depicted in Figure 1 below.

Figure 1: The scree plot of the 2nd PCA

As observed in Figure 1, there was a clear break after the sixth component, not after the eleventh one. In order to find more evidence to determine the number of factors, the data was analyzed via Parallel Analysis using MonteCarloPA.exe, bearing in mind that the obtained eigenvalues for each component should be larger than its corresponding random eigenvalue. Table 4 presents the results.

Table 4: The Obtained and Random Eigenvalues Achieved through Parallel Analysis

Components	Random Eigenvalue	Obtained eigenvalues
1	2.32	24.31
2	2.22	12.81
3	2.16	11.35
4	2.15	9.62
5	2.14	8.90
6	2.13	4.66
7	2.12	2.11
8	1.92	1.86
9	1.88	1.23
10	1.84	1.14
11	1.81	1.00

Table 4 shows that the criterion was met for components 1 to 6. Therefore, based on the low variance that these factors can explain, scree plot, and parallel analysis, it was decided to keep the first six components for further investigation. Table 5 indicates the factor loadings for the items using Promax rotation, which were all greater than .3.

As can be seen in Table 5, factor 1 (method) includes items 35, 39, 40, 47, 48, 49, 50, 59, 63, 64, 66, 67, 68, 69, 70, 76, 81, 82, 83, 84, 85, 100, and 101, (item 70 loaded on both factor 1 and 11; however, its loading on factor 1 is greater, and factor 11 was decided to be excluded), factor 2 entails (outcome) items 13, 18, 19, 72, 86, 90, 92, 95, 98, 105, 91, 93, 99, 103, 104, 12, and 15, factor 3 embodies (perception) items 1, 2, 3, 6, 8, 29, 31, 34, 52, 55, 56, 58, 60, 61, 65, 94, 38, 62, factor 4 consists of (purpose) items 7, 14, 16, 53, 74, 75, 80, 87, 89, 96, 102, 9, and 10, factor 5 comprises (content) items 22, 23, 30, 37, 57, 79, 88, 21, 26, 27, and 28, and factor 6 (system) includes items 32, 33, 51, 71, 73, 77, 97, and 36 (item 32 and 33 loaded on both factor 6 and 11; however, their loadings on factor 6 were greater, and factor 11 was decided to be excluded). Item 11 and 46 loaded on factor 7, item 20 and 4 on factor 8, item 43 and 78 on factor 9, item 45 and 5 on factor 10, and item 41 and 42 on factor 11. According to Pallant (2013), more than two items must be loaded on each factor; as a result, these items which apparently measured some other constructs were excluded from further analysis. After the omission of these ten items (4, 5, 11, 20, 41, 42, 43, 45, 46, and 78), another PCA (the 3rd PCA) was run. In the 3rd PCA, the significant level of Bartlett's test of sphericity was greater than .001 and Kaiser-Meyer-Olkin value was .92, both of which met the criteria of factorability. Table 6 indicated the results of the factors with the eigenvalues of more than 1.

Table 5: Pattern Matrix for 100 Items in the second PCA

	Component										
	1	2	3	4	5	6	7	8	9	10	11
I59	.93										
I40	.93										
I48	.92										
I6	.92										
I101	.91										
I84	.91										
I100	.90										
I66	.90										
I63	.90										
I67	.89										
I83	.89										
I69	.88										
I50	.87										
I47	.87										
I85	.87										
I82	.86										
I81	.85										
I64	.82										
I68	.82										
I49	.80										
I76	.80										
I39	.76										
I35	.70										.355
I70	.67										
I18		.96									
I12		.94									
I91		.93									
I13		.92									
I104		.90									
I93		.90									
I99		.9									
I86		.90									
I105		.88									
I98		.88									
I103		.87									
I15		.87									
I90		.87									
I19		.87									
I92		.86									
I72		.80									
I95		.80									
I94			.91								
I38			.90								
I31			.89								
I55			.89								
I65			.89								
I61			.86								
I56			.86								
I1			.85								

I60	.85								
I3	.85								
I52	.84								
I62	.84								
I29	.84								
I2	.82								
I34	.78								
I58	.78								
I8	.77								
I87		.94							
I102		.93							
I89		.92							
I80		.92							
I74		.92							
I10		.92							
I14		.91							
I96		.89							
I53		.89							
I7		.85							
I75		.84							
I9		.80							
I16		.80							
I28			.90						
I27			.90						
I37			.89						
I23			.86						
I22			.86						
I57			.86						
I79			.85						
I21			.83						
I88			.83						
I26			.80						
I30			.79						
I33				.87					-.32
I97				.85					
I71				.85					
I77				.81					
I51				.80					
I73				.76					
I36				.76					
I32				.66					.30
I46					.93				
I11					.93				
I4						.96			
I20						.95			
I78							.78		
I43							.63		
I45								.79	
I5								.75	
I42									.82
I41									.56

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

Table 6: The Results for the Factors with the Eigenvalues of more than 1 in the 3rd PCA

	Initial Eigenvalues			**Extraction Sums of Squared Loadings**			**Rotation Sums of Squared Loadings**[a]
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**
1	22.9	25.51	25.51	22.9	25.51	25.51	20.14
2	6	14.14	39.65	6	14.14	39.65	14.46
3	12.7	12.48	52.13	12.7	12.48	52.13	13.71
4	3	10.61	62.74	3	10.61	62.74	11.74
5	11.2	8.69	71.44	11.2	8.69	71.44	10.17
6	3	4.95	76.40	3	4.95	76.40	7.97
	9.54			9.54			
	7.82			7.82			
	4.46			4.46			

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

As seen in Table 6, there were six components with eigenvalues of more than 1, which explained a total of 76.40 percent of the variance. The first, second, third, fourth, fifth, and sixth factors could explain almost 26%, 14%, 12%, 11%, 9%, and 5% of the total variance respectively. The scree plot of the data is presented in Figure 2 below.

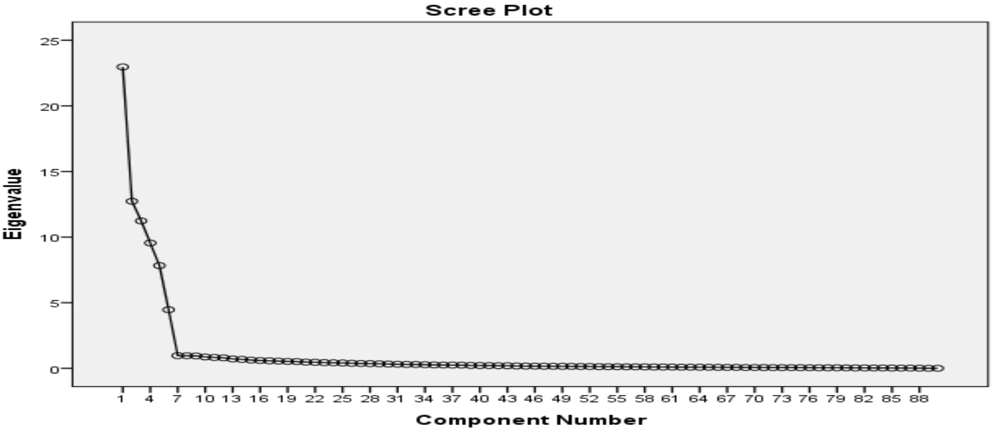


Figure 2: The scree plot in the 3rd PCA

As shown in Figure 2, there is a clear break after the sixth component. Table 7 indicates the factor loadings for the items using Promax rotation, which were all greater than .3.

Table 7: Pattern Matrix for 90 Items in the 3rd PCA

	Component					
	1	2	3	4	5	6
I100	.93					
I66	.92					
I40	.91					
I84	.91					
I6	.90					
I101	.90					
I81	.90					
I48	.89					
I50	.89					
I47	.88					
I59	.88					
I83	.88					
I69	.87					
I63	.86					
I76	.86					
I39	.85					
I85	.85					
I67	.85					
I82	.83					

I68	.83			
I49	.82			
I35	.82			
I64	.79			
I70	.71			
I18		.96		
I12		.94		
I91		.93		
I13		.92		
I93		.90		
I104		.90		
I86		.90		
I19		.89		
I15		.89		
I99		.88		
I105		.88		
I103		.88		
I92		.88		
I90		.88		
I98		.87		
I72		.80		
I95		.79		
I94			.89	
I3			.89	
I1			.89	
I62			.88	
I55			.88	
I38			.87	
I65			.87	
I31			.87	
I2			.86	
I60			.85	
I56			.84	
I29			.83	
I8			.82	
I34			.82	
I61			.82	
I52			.80	
I58			.75	
I14				.92
I87				.92
I102				.92
I74				.91
I80				.91

I96				.91		
I7				.90		
I89				.90		
I53				.89		
I10				.89		
I16				.86		
I75				.82		
I9				.82		
I23					.94	
I22					.93	
I21					.91	
I27					.88	
I79					.87	
I26					.86	
I37					.85	
I28					.85	
I30					.84	
I88					.84	
I57					.79	
I97						.90
I71						.90
I51						.81
I36						.81
I32						.79
I73						.78
I77						.74
I33						.70

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

The results also indicated that there was a weak correlation among the six components, as shown in Table 8.

Table 8: Component Correlation Matrix of the Six Extracted Components

Component	**1**	**2**	**3**	**4**	**5**	**6**
1	1.00	.13	.19	.19	.19	.22
2	.13	1.00	.11	.08	.12	.11
3	.19	.11	1.00	.04	.12	.13
4	.19	.08	.04	1.00	.07	.17
5	.19	.12	.12	.07	1.00	.18
6	.22	.11	.13	.17	.18	1.00

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

As is clear from Table 8, the correlations among factors were very low (the maximum correlation is .22, which is less than .3), which is satisfactory.

Internal consistencies for the whole questionnaire and for the individual extracted factors were calculated through Cronbach's alpha. As a guideline, "measures higher than .7 are considered as acceptable, while measures below .6 are considered as weak to unacceptable" (Dörnyei, 2003, p. 112). As mentioned before, the index for the whole TE questionnaire was .92, showing a high internal consistency. Table 9 shows the results for the individual extracted factors.

Table 9: The Reliability of Individual Factors

Factors	Cronbach's Alpha	N of Items
Factor1	.78	23
Factor2	.83	17
Factor3	.76	18
Factor4	.81	13
Factor5	.79	11
Factor6	.85	8

As seen in Table 9, all of the indices for the internal consistency of the factors were above .7, which indicates an acceptable level of internal consistency.

Confirmatory Factor Analysis (CFA)

Based on the EFA conducted earlier, a six-factor model encompassing method, outcome, perception, purpose, content, and system of teacher evaluation was hypothesized. Testing this hypothesized model, confirmatory factor analysis of the data from the participants was conducted through the AMOS (Arbuckle, 2013). To this end, the revised questionnaire (Appendix B) was administered to 360 more supervisors who were available at that time. To measure the data fit, Incremental Fit Index (IFI), Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMR), Browne-Cudeck Criterion (BCC), Akaike Information Criterion (AIC), Bayes Informatin Criterion (BIC), and Root Mean Square Error of Approximation (RMSEA) were checked. According to Byrne

(2001), CFI is similar to Normed Fit Index (NFI) and is a better measure; therefore, just CFI is reported. In order to see if competing models would better fit the data than the hypothesized model, nested model comparisons were also run. Being based on the classical test theory, these models were congeneric as the hypothesized model, tau-equivalent model, and parallel model (Millsap & Everson, 1991). In the tau-equivalent model, all observed variables have equal factor loadings with their own unique variances. In the parallel model, all observed variables are set to have equal factor loadings and equal unique variances; therefore, the result is the covariance structure. Table 10 indicates the results of all the hypothesized congeneric, tau-equivalent, and parallel models.

Table 10: Goodness of Fit for Three Nested Models

Models	**CMIN**	**Df**	**p**	**CMIN/DF**	**RMSEA**	**PCLOSE**	**SRMR**	**IFI**	**TLI**	**CFI**	**AIC**	**BCC**	**BIC**
Congeneric	5506.79	3900	.000	1.41	.03	1.00	.04	.91	.91	.91	5896.79	5829.58	6654.58
Tau-equivalent	6406.53	3985	.000	1.60	.04	1.00	.05	.87	.87	.87	6626.53	7054.01	7164.01
Parallel	8838.15	4068	.000	2.17	.05	.00	.06	.75	.75	.75	8892.15	8997.07	9024.07

Based on Arbuckle (2013), the chi-square value is the extent to which the data does not fit the estimated model; therefore, the lower its value, the better. As Table 10 indicates, the chi-square value for the congeneric model is the lowest one ($\chi2$=5506.79) with the lowest degree of freedom (*df*=3900) and the p value was .00. Although the p value was significant (which should be non-significant), the ratio of $\chi2$/df for all models was less than 3, so, the results indicated a good fit for all the nested models (Kline, 1994). However, the lower this ratio, the better; as a result, the congeneric model (i.e., 1.41) was the best fit compared to the other two nested models. The RMSEA values here indicated that the data fit well for the congeneric and tau-equivalent models since the values in these models were less than .05 but this value for the parallel model was .05. Moreover, the value of RMSEA for the congeneric model was the lowest (i.e., .03) and the lower the value of RMSEA, the better the data fits. PCLOSE must be above .05, which was the case with the congeneric and tau-equivalent models (i.e., 1.00), but this value was less than .05 for the parallel model. Regarding the SRMR, according to the guidelines for interpreting the output (Arbuckle, 2013; Byrne, 2001), the smaller the SRMR, the better, and the value less than .05 is good. The value of SRMR was less than .05 for the congeneric model (i.e., .04), was equal to .05 for the tau-equivalent model, and was greater than .05 for the parallel model.

Therefore, regarding SRMR, the parallel model was not a good fit, and since the SMRM value for the congeneric model was the lowest, this model indicated a better fit. The more the values of IFI, TLI, and CFI, the better. Considering their range from 0 to 1, the values closer to 1 and above .9 are better fit. The congeneric model indicted the best fit since these values for this model were all closer to 1 and above .9 (*IFI*= .91, *TLI* =.91, and *CFI* =.91). However, these values for the tau-equivalent and parallel models were less than .9; so, these models did not show good fit. Regarding the values of AIC, BCC, and BIC, the lower values indicate a better fit model and large values show bad fit. As seen in Table 10, the lowest value belonged to the congeneric model (*AIC*=5896.79, *BCC*=5829.58, and BIC=6654.58) compared to other nested models. These results suggested that the congeneric model was a better model fit.

In order to examine whether the tau-equivalent and parallel models fit as much as the congeneric model does, AMOS generates chi-square tests. Table 11 indicates the results of the comparison between the tau-equivalent and congeneric models and between the parallel and congeneric models while considering the congeneric model (the most unrestricted model) correct.

Table 11: The Statistical Comparison among Nested Models

Part A: Assuming model Congeneric to be correct:

Model				**NFI**	**IFI**	**RFI**	
	DF	**CMIN**	**P**			**rho-**	**TLI**
				Delta-1	**Delta-2**	**1**	
							rho2
Tau-equivalent	85	899.74	.00	.03	.04	.03	.04
Parallel	168	3331.35	.00	.14	.17	.13	.15

Based on Table 11, it can be stated that the difference in the chi-squares of the tau-equivalent and congeneric models (=899.74) was significant (p<.05), and the difference in the chi-squares of the parallel and congeneric model (=3331.35) was also significant (*p*<.05). These results indicated that the congeneric model significantly fit the data better. Based on what was stated, it can be declared that the hypothesized model fit the data better than the other two models, and our hypothesized model was confirmed. The CFA of the hypothesized model is presented in Appendix C. In the diagram path, the standardized factor loadings were all above .3, which shows a satisfactory level of factor loadings.

DISCUSSION AND CONCLUSION

This study aimed to explore the perceptions of the evaluators, the criteria, the methods and systems they had in evaluating English language teachers' performance and to develop a reliable and valid teacher evaluation questionnaire. Based on the results, six major factors were reflected by the evaluators, including the method, system, content, perception, purpose, and outcome of evaluation. This finding aligns with various studies reporting planning and purpose (Bryant et al., 2016), classroom observation and learning environment (Ruprich & Urhahne, 2015), and teacher effectiveness and TE methods (Martinez et al., 2016) as the subscales of TE. This indicates that evaluators share some common perceptions about TE probably because certain principles remain the same despite the varieties in educational contexts and cultures.

From among TE constructs, based on the results, most evaluators argued for teacher development. This is congruent with Wang and Day's (2002) opinion that supervisors should shift their perceptions of classroom observations from a means of teacher evaluation to a tool to promote teacher development. As for the methods of evaluation, some participants voted for self-assessment and reflection by teachers (Iyer-O'Sullivan, 2015; Shoffner, 2009; Tripp, 2012), classroom observation followed by post-observation debriefing session (Engin, 2015; Iyer-O'Sullivan, 2015; Mercado & Mann, 2015), and peer observation (Hawkins & Irujo, 2004; Oprandy, 2002).

The analysis of the systems of evaluation showed that most of the teacher evaluations in the context of this study have been summative which could be due to many reasons including the difficulties of doing formative evaluation and the resistance of the evaluators. The results emphasized using formative evaluation along with summative evaluation in teacher evaluation systems (Howard, 2015; King, 2015). With regard to the content of evaluation, the main items included teaching the subject matter and language skills (Howard, 2015), formal and informal chats between teachers and evaluators (Rivera, 2011), the debriefing sessions between teachers and evaluators (Atkinson & Bolt, 2010), corrective feedback (King, 2015; Randall, 2015), and medium of transmission (Freeman, Orzulak, & Morrisey, 2009).

As regards the purpose of evaluation, most of the participants stated the main purpose is to motivate and help teachers (King, 2015). However, some other purposes have been stated such as decision making on remuneration and contracts (Riera, 2011), teacher growth (Clenchy, 2017; DeMatthews, 2015), and improve student performance and learning (Beare, 1989). The reason for this variety stems from the context-dependency nature of the TE purposes. Considering the outcome, the results of this study are consistent with other studies focusing on decisions such as retention, reward, and change in salary (Bello & Jakada, 2017; Ingvarson, Kleinhenz, & Wilkinson, 2007; Odden, Kelley, Heneman, & Milanowski, 2001) and a variety of non-monetary rewards including job promotion and public recognition (Tumaini, 2015).

Furthermore, the results revealed the participants' preferences to have closer relationships with the teachers (Mann & Walsh, 2013), show trust and respect (Clenchy, 2017), share power with teachers (Mercado & Mann, 2015), and assist them in their professional development (Hobson, Ashby, Malderez, & Tomlinson, 2009; Kwan & López-Real, 2010). Regarding the criteria for effective teacher performance, in the present study, some evaluation criteria reflected by the participants involved students' test scores, teachers' classroom management, teachers' professional development, use of humor in the class, and students' and parents' feedback. However, in the literature, some other criteria have been proposed such as teachers being models (Kennedy, 2010), student-centered classrooms (Hunt, 2015), and classroom decision-making (Mercado & Mann, 2015). Thus, the diversity in criteria for evaluating teachers indicates that teaching is a complicated activity which depends on specific context and audience (Quirke, 2015). Razmjoo and Mavaddat (2016) also believe that teacher evaluation and the way it is done have a direct impact on the teacher's performance.

The questionnaire developed in this research surpassed the underpinning theoretical background of Martinez et al.'s (2016) study as it incorporated new themes to TE including perception, purpose, and content of evaluation. As Kane, Kerr, and Pianta (2014) put it, "there is no shortage of debate and opinion on the challenges and promises of teacher performance evaluation, with interests weighing in on all sides" (p. 583). After ensuring its reliability, the results of EFA and CFA analyses indicated that the TE questionnaire developed and validated has good psychometric properties (i.e., construct validity). Thus, the TE questionnaire can be applied for both research and pedagogy. For research, a reliable and valid measure of TE can set teacher evaluators free of evaluating teachers on subjective criteria and/or mere observation of teacher performance in the classroom. By the same token, it avoids evaluators from implementing self-made questionnaires and checklists the validity and reliability of which have not been well-established. For pedagogy, the results of this study would be highly useful in representing the inventories of evaluators' various perceptions and practices, especially when institutional constraints require that evaluators work as per rules and obligations defined by the institute during their course of evaluation. By understanding the principles and procedures of teacher evaluation, the teachers can make informed choices regarding their pedagogical behavior and practices.

Like any other research study, the present study suffered from certain limitations which should be kept in mind. First, the social, cultural, academic, ethnic, cognitive, and emotional backgrounds of the participants of the study constituted the primary limitation which could not be truly controlled. Second, many language institutes were not cooperative because they did not like to be criticized for their teacher evaluation system or lose their reputation. Third, contrary to the number of teachers teaching in English language institutes, the number of observers and evaluators in each institute was limited to a few numbers. Thus, as far as generalizability is concerned, wider application of the tested model and the developed TE questionnaire is required. In other words, cross-validation of the TE questionnaire can be carried out with teacher evaluators within as wide a number of EFL contexts as possible to be able to make claims for the generalizability of the tested model and wider application of the developed TE questionnaire.

Validating a data collection instrument (e.g., a questionnaire) is a cyclical process which does not stop even after the instrument has been initially validated. Therefore, replication studies are required that provide further validation from several dimensions. Both convergent and divergent validation studies are recommended, using insightful theories and models of teacher evaluation. In addition, researchers involved in the field of teacher evaluation can replicate this study using other data collection methods such as classroom observations, interviews, document reviews, surveys, and recording of debriefing sessions which can lead them to a more comprehensive understanding of teacher evaluation. Finally, longitudinal studies can be conducted to examine how the changes in the evaluators' perceptions would affect teachers and their performance in the classroom.

References

Arbuckle, J. L. (2013). *IBM SPSS Amos 22 user's guide*. Crawfordville, FL: IBM Corp.

Atkinson, D. J., & Bolt, S. (2010). Using teaching observations to reflect upon and improve teaching practice in higher education. *Journal of the Scholarship of Teaching and Learning*, *10*(3), 1-19.

Beare, H. (1989). The Australian policy context. In J. Lokan & P. McKenzie (Eds.), *Teacher appraisal: Issues and approaches* (pp. 5-8). Victoria, Australia: Australian Council for Educational Research.

Behlow, D. J. (1990). *The Purposes, procedures, and outcomes of evaluating Wisconsin public school principals.* Madison: Wisconsin University.

Bello, B., & Jakada, B. (2017). Monetary reward and teachers' performance in selected public secondary schools in Kano state. *Journal of Education and Practice 8*(7), 1-14.

Bryant, C. L., Maarouf, S., Burcham, J. G., & Greer, D. (2016). The examination of a teacher candidate assessment rubric: A confirmatory factor analysis. *Teaching and Teacher Education, 57*(2), 79-96. https://doi.org/10.1016/j.tate.2016.03.012

Byrne, B. M. (2001). *Structural equation modelling with Amos, basic concepts, applications, and programming*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Chalmers, D., & Gardiner, D. (2015). An evaluation framework for identifying the effectiveness and impact of academic teacher development programs. *Studies in Educational Evaluation*, *46*(4), 81-91. https://doi.org/10.1016/j.stueduc.2015.02.002

Clenchy, K. R. (2017). *Teacher evaluation models: Compliance or growth oriented?* (Doctoral dissertation). Retrieved from https://repository.library.northeastern.edu/files/neu:cj82qm547/fulltext.pdf

Danielson, C. (2013). *The framework for teaching evaluation instrument*. Princeton, NJ: The Danielson Group.

Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching and Teacher Education, 36*(4), 1-11. https://doi.org/10.1016/j.tate.2013.06.011

DeMatthews, D. (2015). Principal and teacher collaboration: An exploration of distributed leadership in professional learning communities. *International Journal of Educational Leadership and Management, 2*(2), 176-206.

Dörnyei, Z. (2003). *Questionnaires in second language research. Construction, administration, and processing*. London: Lawrence Erlbaum Associates Publishers.

Engin, M. (2015). Artefacts in scaffolding the construction of teaching knowledge. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 85-99). London: Bloomsbury.

Freeman, D., Orzulak, M. M., & Morrisey, G. (2009). Assessment in second language teacher education. In A. Burns & J. C. Richards (Eds.), *The Cambridge guide to second language teacher education.* (pp. 77-90). Cambridge: Cambridge University Press.

Fyson, N. (1993). *The evaluation of secondary school principals as perceived by principals and district level administrators* (Doctoral dissertation). University of Southern California, California.

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning. Research & policy brief.* Washington, DC: National Comprehensive Center for Teacher Quality.

Hawkins, M., & Irujo, S. (2004). *Collaborative conversations among language teacher educators*. Alexandria, VA: TESOL.

Hobson, A. J., Ashby, P., Malderez, A., & Tomlinson, P. D. (2009). Mentoring beginning teachers: What we know and what we don't. *Teaching and Teacher Education, 25* (1), 207-216. https://doi.org/10.1016/j.tate.2008.09.001

Howard, A. (2010). *Teacher appraisal: The impact of observation on teachers' classroom behavior* (Doctoral dissertation). The University of Warwick, Warwick, England.

Howard, A. (2015). Giving voice to participants in second language education evaluation. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 193-210). London: Bloomsbury.

Hunt, N. (2015). Student teacher placements: A critical commentary. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 151-164). London: Bloomsbury.

Ingvarson, L., Kleinhenz, E., & Wilkinson, J. (2007). *Research on performance pay for teachers.* Melbourne, Vic, Australia: Australian Council for Educational Research (ACER), Retrieved from
http://research.acer.edu.au/cgi/viewcontent.cgi?article=1000&context=workforce

Iyer-O'Sullivan, R. (2015). From bit to whole: Reframing feedback dialogue through critical incidents. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 69-84). London: Bloomsbury.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project.* San Francisco, CA: Jossey-Bass, a Wiley brand.

Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching.* Seattle, WA: The Bill and Melinda Gates Foundation.

Kennedy, M. (2010). *Teacher assessment and the quest for teacher quality.* San Francisco, CA: Jossey-Bass.

King, M. (2015). Evaluating experienced teachers. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 167-179). London: Bloomsbury.

Kline, P. (1994). *An Easy guide to factor analysis*. London: Routledge.

Kwan, T., & López-Real, F. (2010). Identity formation of teacher-mentors: An analysis of contrasting experiences using a Wengerian matrix framework. *Teaching and Teacher Education, 26*(3), 722-731. https://doi.org/10.1016/j.tate.2009.10.008

Kyriakides, L., Campbell, R. J., & Christofidou, E. (2002). Generating criteria for measuring teacher effectiveness through a self-evaluation approach: A complementary way of measuring teacher effectiveness. *School Effectiveness and School Improvement, 13*(3), 291-325. https://doi.org/10.1076/sesi.13.3.291.3426

Kyriakides, L., Demetriou, D., & Charalambous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research, 48*(1), 1-20. https://doi.org/10.1080/j.tate.2008.09.001

Lowe, A. M. (2000). *A study of the evaluation of secondary school teachers in selected schools in southern California as perceived by secondary school teachers and evaluators* (Unpublished doctoral dissertation). AZUSA Pacific University, California.

Mann, S., & Walsh, S. (2013). RP or RIP: A critical perspective on reflective practice. *Applied Linguistics Review Journal, 4*(2), 291-315. https://doi.org/10.1515/applirev.2013.07.003

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*(2), 15-29. https://doi.org/10.1016/j.stueduc.2016.03.002

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures, and software solution*. Austria: Klagenfurt.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based evaluation of teacher performance: An empirical approach.* New York, NY: Longman.

Mercado, L. A., & Mann, S. (2015). Mentoring for teacher evaluation and development. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 35-54). London: Bloomsbury.

Millsap, R. E. & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioural Research, 26*(3), 479-497. https://doi.org/10.1207/j.tate.2009.09.002

Moradi, Kh., Sepehrifar, S., & Parhizkar Khadiv, T. (2014). Exploring Iranian EFL teachers' perceptions on supervision. *Procedia - Social and Behavioral Sciences, 98*(6), 1214-1223. https://doi.org/10.1016/j.sbspro.2014.03.536

Odden, A., Kelly, C., Heneman, H., & Milanowski, A. (2001). Enhancing teacher quality through knowledge-and skills-based pay. *Journal of Policy Briefs, 18*(1), 51-71. https://doi.org/10.1037/ j.tate.2012.10.001

Oprandy, B. (2002). A counseling-learning perspective. In J. Edge (Ed.), *Continuing professional development* (pp. 252-264). UK: International Association of Teachers of English as a Foreign Language (IATEFL).

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (4th Ed.). Crows Nest, NSW: Allen & Unwin.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd Ed.). Thousand Oaks, CA: Corwin press.

Prasertsin, U. (2015). Confirmatory factor analysis of teacher's work for integrating research, evaluation measurement and quality assurance model. *Social and Behavioural Sciences, 197*(1), 2201-2206. https://doi.org/10.1016/j.sbspro.2015.07.007

Quirke, P. (2015). A system for teacher evaluation. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 101-114). London, England: Bloomsbury.

Rahmany, R., Hasani, M. T., & Parhoodeh, K. (2014). EFL teachers' attitudes towards being supervised in an EFL context. *Journal of*

Language Teaching and Research, 5(2), 348-359. https://doi.org/10.4304/jltr.2014.09.008

Randall, M. (2015). Observing for feedback: A counselling perspective. In A. Howard & H. Donaghue (Eds.), *Teacher evaluation in second language education* (pp. 55-65). London: Bloomsbury.

Razmjoo, S. A., & Mavaddat, R. (2016). Understanding professional challenges faced by Iranian teachers of English. *International Journal of English Linguistics,* *6*(3), 208-220. https://doi.org/10.5539/ijel.v6n3p208

Riera, G. (2011). New directions in teacher appraisal and development. In C. Coombe L. Stephenson & S. Abu-Rmaileh (Eds.), *Leadership and management in English language teaching* (pp. 49-66). Dubai: TESOL Arabia.

Rivera, D. H. (2011). *How to encourage informal debriefing? A step further on changing attitudes with games* (Master's thesis). Retrieved from
http://www.diva-ortal.org/smash/get/diva2:426448/FULLTEXT01.pdf

Ruprich, C., & Urhahne, D. (2015). Development of a questionnaire for the assessment of teacher goals from a content perspective. *International Journal of Educational Research,* *72*(2), 173-184. https://doi.org/10.1016/j.ijer.2015.06.005

Shoffner, M. (2009). The place of the personal: Exploring the affective domain through reflection in teacher preparation. *Teaching and Teacher Education, 25*(6), 783-789.
https://doi.org/10.1016/j.tate.2008.11.012

Stufflebeam, D. L. (1969). Evaluation as enlightenment for decision-making. In A. Walcott (Ed.), *Improving educational assessment and an inventory of measures of affective behavior* (pp. 41-73). Washington, DC: Association for supervision and curriculum development.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics.* Boston: Allyn & Bacon.

Tripp, D. (2012). *Critical incidents in teaching: Developing professional judgement.* London: Routledge.

Tumaini, M. (2015). *The contribution of non-monetary incentives to teachers' retention in public secondary schools in Korogwe urban* (Doctoral dissertation). Retrieved from http://repository.out.ac.tz/1416/1/mary_submission.pdf

Wang, W., & Day, C. (2002). *Issues and concerns about classroom observation: Teachers' perspectives*. Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (TESOL), St. Louis, MO, USA.

Appendix A

Teacher Evaluator Interview

1. How do you define teacher evaluation generally?
2. Which do you prefer? Being an evaluator or an evaluated teacher? Why?
3. What factors do you consider in evaluating EFL teachers?
4. What procedures do you follow in order to evaluate EFL teachers?
5. How often do you evaluate teachers in one semester?
6. Are the evaluator-teacher debriefing sessions held formally or informally?
7. Do teachers actually behave as instructed in the debriefing sessions?
8. Do you have freedom in the process of teacher evaluation or do you thoroughly follow the predetermined rules of your institute?
9. Which one do you prefer? Formative or summative evaluation? Why?
10. In what ways do you think evaluation should influence your personal growth and professional development?

Appendix B

Teacher Evaluators' Perception Questionnaire

Identifying the perceptions and evaluation systems of EFL Teacher Evaluators is considered an essential endeavor for designing and implementing effective teacher evaluation programs within the realm of teacher education. The aim of this questionnaire is to explore the perceptions of Iranian EFL teacher evaluators (supervisors) regarding the teacher evaluation systems. We assure you of the confidentiality of your responses.

Section I: Demographic Information

Gender: Male: Female: Age:
City:
Your primary role in teacher evaluation is: Supervising Professor Teacher Mentor
Grades you teach/supervise:
Number of years you have served in your teacher evaluation role:
Less than 1 1-3 4-6 7-10 Over 10

Section II: Please indicate how important these qualities are by choosing the relevant number on a scale of 1 to 5:

1 = *Totally Disagree;*

2 = *Disagree*;
3 = *No Idea*;
3 = *Somehow Agree;*
5 = *Totally Agree.*

The Current Teacher Evaluation (TE):	**1 Totally Disagree**	**2 Disagree**	**3 No Idea**	**4 Somehow Agree**	**5 Totally Agree**
1. is necessary for teachers	1	2	3	4	5
2. identifies and rewards outstanding teachers	1	2	3	4	5
3. identifies and terminates incompetent teachers	1	2	3	4	5
4. is authoritative rather than democratic	1	2	3	4	5
5. helps teachers to overcome instructional problems	1	2	3	4	5
6. is inspectional rather than a collaborative process	1	2	3	4	5
7. aims to control rather than improve	1	2	3	4	5
8. encourages teachers toward better performance	1	2	3	4	5
9. contributes to the personal growth of teachers	1	2	3	4	5
10. contributes to the teachers' professional development	1	2	3	4	5
11. increases teachers' knowledge of teaching methodologies	1	2	3	4	5
12. assists teachers in decision-making	1	2	3	4	5
13. improves teachers' teaching skills and students' achievement	1	2	3	4	5
14. provides constructive feedback to the teachers	1	2	3	4	5
15. promotes the reputation of the language institute	1	2	3	4	5

Teacher Evaluation (TE) is Based on:					
16. teacher's skill in planning	1	2	3	4	5
17. teacher's skill in assessment and evaluation	1	2	3	4	5
18. teacher's use of instructional materials	1	2	3	4	5
19. teacher's ability to recognize and provide for individual differences	1	2	3	4	5
20. teacher's oral and written communication skills	1	2	3	4	5
21. teacher's classroom routines and control	1	2	3	4	5
22. teacher's fairness in dealing with students	1	2	3	4	5
23. teacher's knowledge of the subject matter	1	2	3	4	5
24. teacher's cooperative approach toward parents and school personnel	1	2	3	4	5
25. teacher's adherence to school policies/procedures	1	2	3	4	5
In the Current Teacher Evaluation (TE):					
26. the regulations of the institute and the dominant teacher evaluation system at the work place are strictly followed	1	2	3	4	5
27. teachers openly accept the criticisms that evaluators point out	1	2	3	4	5
28. the entire process of TE is described to the teachers	1	2	3	4	5
29. responsibilities are jointly shared between the teacher and the evaluator	1	2	3	4	5
30. class management and overcoming students' behavioral problems are	1	2	3	4	5

considered					
31. promoting student's positive attitudes toward learning English is significant	1	2	3	4	5
Teacher Evaluation (TE) is Done through:					
32. pre-evaluation conference	1	2	3	4	5
33. post-evaluation conference	1	2	3	4	5
34. feedback by checklist or rating scale	1	2	3	4	5
35. classroom observation by an administrator	1	2	3	4	5
36. classroom observation by another teacher	1	2	3	4	5
37. classroom observation by a mentor teacher	1	2	3	4	5
38. recommendation by a teacher consultant regarding the renewal of the contract of another teacher	1	2	3	4	5
As an Evaluator					
39. I prefer to be an evaluator rather than a teacher.	1	2	3	4	5
40. I exactly know the purpose of teacher evaluation I do.	1	2	3	4	5
41. I think the method of teaching is the most important factor in evaluation.	1	2	3	4	5
42. I pay attention to teacher's personality in evaluation.	1	2	3	4	5
43. I believe the teacher's English proficiency can compensate for other shortcomings.	1	2	3	4	5
44. I believe teacher-student rapport determines teacher's	1	2	3	4	5

future job status.					
45. I think teachers should act just like ordinary days while being observed.	1	2	3	4	5
46. I suppose supervision negatively affects teacher's performance.	1	2	3	4	5
47. I think focusing on positive points is under the shadow of negative points.	1	2	3	4	5
48. I observe teachers without any interruptions.	1	2	3	4	5
49. I merely take notes of teacher's mistakes and errors while making observation.	1	2	3	4	5
50. I believe teacher evaluation is merely based on classroom observation.	1	2	3	4	5
51. I think teacher evaluation should happen in the post-observation debriefing sessions.	1	2	3	4	5
52. I start the debriefing sessions with negative points.	1	2	3	4	5
53. I give the opportunity to the teacher to clarify and explain his/her points.	1	2	3	4	5
54. I postpone my recommendations to the last minutes of the debriefing sessions.	1	2	3	4	5
55. I prefer formative evaluation rather than summative evaluation.	1	2	3	4	5
56. The kind of evaluation I do is basically summative evaluation.	1	2	3	4	5
57. I believe the current teacher evaluation system is	1	2	3	4	5

working well for language institutes.					
58. I think teachers often benefit from teacher evaluation.	1	2	3	4	5
59. I assume the current TE system gives priority to the evaluators in relation to the teachers.	1	2	3	4	5
60. I believe teachers should be behaved as instructed in the debriefing sessions.	1	2	3	4	5
61. I often behave as arranged in the debriefing sessions.	1	2	3	4	5
62. I think the evaluator-evaluated teachers debriefing sessions should be held formally.	1	2	3	4	5
63. I have freedom to make changes in TE process that the institute has dictated.	1	2	3	4	5
64. I think teachers should be informed about the evaluation criteria before observation.	1	2	3	4	5
65. I seek mutual compromise rather than unilateral recommendations in the debriefing sessions.	1	2	3	4	5
66. I follow pre-observation talk, observation, and post-observation conference procedure.	1	2	3	4	5
An Evaluator					
67. cries out sudden classroom visits to evaluate the teachers' performance.	1	2	3	4	5
68. holds educational workshops to train teachers on operating and the use of	1	2	3	4	5

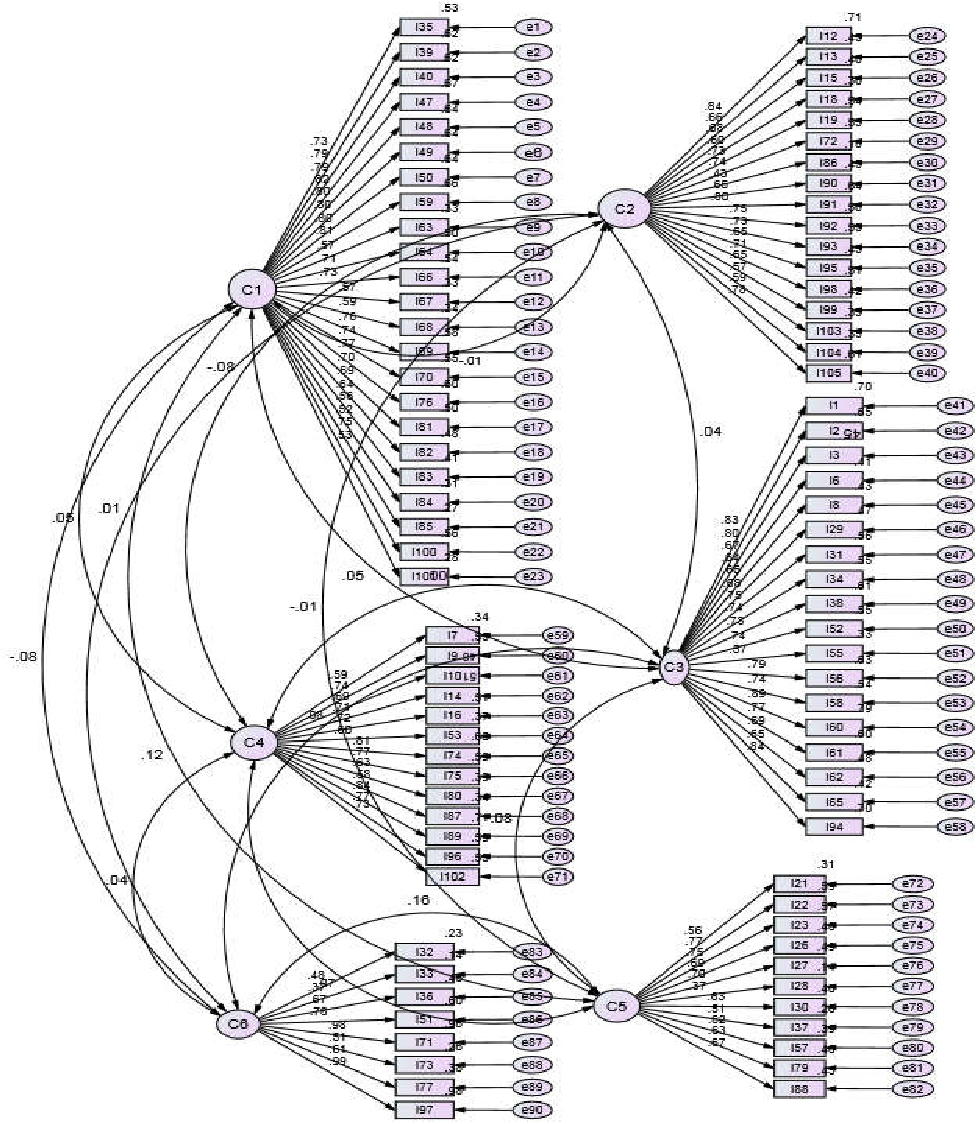
appropriate teaching aids.					
69. holds regular meetings for English teachers to discuss the teaching difficulties and proposes the methods of overcoming them.	1	2	3	4	5
70. makes critical judgments and concentrates on teachers' mistakes.	1	2	3	4	5
71. clarifies to the teachers the importance of feedback and reinforcement in the teaching process.	1	2	3	4	5
72. prompts the teachers' positive attitude toward teaching English.	1	2	3	4	5
73. helps the teachers to develop their ability to speak correct and fluent English.	1	2	3	4	5
74. Works with teachers to determine the students' needs.	1	2	3	4	5
Teacher Evaluation					
75. reinforces the strengths of a teacher.	1	2	3	4	5
76. makes improvements in teaching and teacher's skills.	1	2	3	4	5
77. assists in making personnel decisions related to promotion or termination.	1	2	3	4	5
78. rewards a teacher for excellence.	1	2	3	4	5
79. accurately reflects job performance.	1	2	3	4	5
80. is a formality without consequences in the improvement of the teacher practice.	1	2	3	4	5
81. is merely a controlling instrument for the teachers'	1	2	3	4	5

performance.					
82. makes the relationship between the evaluator and the teacher to be hierarchical.	1	2	3	4	5
83. leads to value judgments about the overall quality of teacher competences.	1	2	3	4	5
84. fosters the culture of competitiveness among teachers.	1	2	3	4	5
85. makes use of data collection based on individual criteria.	1	2	3	4	5
86. makes use of data collection based on standardized criteria.	1	2	3	4	5
87. is merely a bureaucratic ritual.	1	2	3	4	5
88. promotes collaborative work.	1	2	3	4	5
89. strengthens the school's professional climate.	1	2	3	4	5
90. aids in the improvement of the educational program.	1	2	3	4	5

Constructs of the Questionnaire

1. Purposes for Evaluating Teachers
2. Content for Evaluating Teachers
3. Teacher Evaluation Systems
4. Method of Evaluating Teachers
5. Perception and Procedure for Teacher Evaluators
6. Outcomes of Evaluating Teachers

Appendix C
The six-factor hypothesized model


C1
C2
C3
C4
C5
C6
I35
I39
I40
I47
I48
I49
I50
I59
I63
I64
I66
I67
I68
I69
I70
I76
I81
I82
I83
I84
I85
I100
I101
I12
I13
I15
I18
I19
I72
I86
I90
I91
I92
I93
I95
I98
I99
I103
I104
I105
I1
I2
I3
I6
I8
I29
I31
I34
I38
I52
I55
I56
I58
I60
I61
I62
I65
I94
I7
I9
I10
I14
I16
I53
I74
I75
I80
I87
I89
I96
I102
I21
I22
I23
I26
I27
I28
I30
I37
I57
I79
I88
I32
I33
I36
I51
I71
I73
I77
I97
e1
e2
e3
e4
e5
e6
e7
e8
e9
e10
e11
e12
e13
e14
e15
e16
e17
e18
e19
e20
e21
e22
e23
e24
e25
e26
e27
e28
e29
e30
e31
e32
e33
e34
e35
e36
e37
e38
e39
e40
e41
e42
e43
e44
e45
e46
e47
e48
e49
e50
e51
e52
e53
e54
e55
e56
e57
e58
e59
e60
e61
e62
e63
e64
e65
e66
e67
e68
e69
e70
e71
e72
e73
e74
e75
e76
e77
e78
e79
e80
e81
e82
e83
e84
e85
e86
e87
e88
e89
e90
.53
.71
.70
.34
.31
.23
-.08
.05
.01
-.08
.04
.05
-.01
.12
.04
.16
-.01
.08