Research in English Language Pedagogy (2022)10(2): 256-276

DOI: 10.30486/RELP.2021.1939981.1306

Original Article

Task-Based Speaking Assessment in an EFL Academic Context: A Case of Summative and Formative Assessment

Reza Bagheri Nevisi[*,1], *Rasoul Mohammad Hosseinpur*[1]

[1] English Language and Literature Department, University of Qom, Qom, Iran

Submission date: 9 Sept, 2021 Acceptance date: 19 Dec, 2021

Abstract

Consistent with the paradigm shift in language assessment from psychometrics to educational assessment, from an examination culture to an assessment culture, the present study was an attempt to compare and contrast the two forms of speaking assessment: Summative and formative and to see how much consistency existed between the two. To this end, 46 undergraduate Iranian EFL students participated in the study. To achieve the formative assessment, some pedagogic speaking tasks were designed and EFL learners' speaking abilities were assessed over a three-month period based on pre-determined criteria. As for the summative assessment, a semi-structured interview was conducted at the end of the course, and learners' performances were assessed by two different raters based on the same criteria. To analyze the data, descriptive statistics, MANOVA, and Pearson correlation were utilized. The results indicated a significant agreement between formative assessment of the first-rater and summative assessment of the second-rater. The findings revealed that from both formative and summative perspectives, pronunciation posed the least challenge whereas coherence and range presented the greatest difficulty to EFL language learners. The study implies that the formative and summative assessment procedures will have to be integrated within classroom settings with more emphasis on the former.

Keywords: Assessment, Coherence, Formative Assessment, Pronunciation, Range, Summative

Corresponding Author's E- mail: re.baghery@gmail.com

1. Introduction

Our personality, self-image, world knowledge, reasoning ability, and the ability to express our thoughts can all be reflected in our spoken performance in a foreign language (Luoma, 2004). Many language learners embark upon the language learning process for the sole reason of being able to communicate well in a second or foreign language context. Accordingly, becoming a competent L2 speaker has become a true objective for a lot of foreign language learners to be properly attained.

To make progress in L2 speaking, teachers need to constantly assess students based on pre-determined criteria and provide suitable feedback accordingly. Assessment forms an integral part of the learning process which must be paid sufficient attention to by educators, curriculum developers, and language practitioners. Throughout the learning process, learners' speaking abilities will have to be assessed by language teachers so that they can be properly provided with the necessary guidance and scaffolding along the way to promote their speaking abilities. There are various specific aims for the assessment including the examination of what pupils have mastered compared to the yardsticks of performance or their mates. Another purpose is to provide teachers and students with appropriate feedback. The feedback can be utilized by educators to rectify their classroom practices, and by language learners to monitor their own progress and make possible amendments in their future performances. A further objective of the assessment is to bring about up-to-the-minute changes and promote developments and innovations in, theory, practice, and policy-making (Brindley, 2013; Brown, 2010, McNamara & Hill, 2011; Shehadeh, 2012).

The terms measurement, test, and evaluation are often wrongly used synonymously and interchangeably (Bachman, 1990). Many make mistakes in using these terms, however, they are quite distinct, each bearing some unique characteristics. Although the definitions put forward for such terminologies in language testing vary to a large extent, they all point to the psychometric nature of testing and the educational spirit inherent in language assessment. For instance, Brown (2010) defines a test as a method of measuring an individual's capability, knowledge, or performance in a given field while Bachman (1990) describes a test as a measurement device designed to obtain a specific sample of an individual's behavior. Furthermore, measurement is described as the process of quantifying the features of individuals based on explicit rules and procedures (Bachman, 2004). In the same vein, evaluation is defined as "the systematic gathering of information to make decisions" Weiss (1972, cited in Bachman, 1990, p.22). Therefore, it can be readily discerned that each term is unique in its own right and has to be applied in a proper context.

Assessment is a popular but often misconstrued term in current pedagogical practices (Brown, 2010). Assessment is a broader term which encompasses both formal and informal measurement tools and other types of qualitative assessment (Chapelle & Brindley, 2010). This study was set up to not only include and incorporate two major types of assessment: Formative and summative in a speaking course but also aims to investigate their probable effectiveness and consistency in an EFL academic context. Moreover, the study also sought to discover the items that posed the maximum and minimum challenge to EFL learners from both summative and formative perspectives.

2. Literature Review

2.1. Views on Assessment

Language assessment has recently undergone a paradigm shift from psychometrics to educational assessment, from a testing and examination culture to an assessment culture. Psychometricians believed in the exact, rigorous, objective, and limiting measurement and psychometric testing is well-rooted in conventional pedagogical models of teaching. This traditional model assumes that knowledge and skill can be decomposed and decontextualized, and this psychometric model of assessment is a static one based on a rigidly defined normal distribution of achievement (Gipps, 1994). Therefore, utilizing the application of nontraditional types of assessment in language classrooms represents the developing paradigm in education in general and second language teaching in particular. In the old paradigm, the focus is mainly on language itself while the new paradigm concentrates on the very notion of communication. The former is seemingly teacher-centered and product-oriented whereas the latter is mostly learner-oriented. The new paradigm also integrates language skills and allows for multiple solutions whereas the old one believes in the isolation of language skills and one-way correctness. Tests that test are typical of the old paradigm while tests that also teach are characteristic of the new paradigm (Richards & Renandya, 2002).

Brown (2010) compared and contrasted the traditional and alternative forms of assessment. Traditional forms of assessment engage language learners in one-shot, standardized exams whereas alternative forms regard assessment as continuous, long-term. Traditional forms of assessment are product-oriented, summative, and time-based which allow for multiple-choice formats, open-ended solutions, and creative answers. However, alternative forms of assessment are process-oriented, formative, and untimed which allow for a free-response format with a focus on the right answer. While decontextualized test items, non-interactive performance, and norm-referenced scorings are typical of the former, the latter focuses on contextualized communicative tasks, interactive performance, and criterion-referenced scorings. Accordingly, it can be said that traditional assessment fosters extrinsic motivation whereas alternative forms of assessment foster intrinsic motivation.

The developers of speaking assessment must have a clear understanding of what speaking is like and then: (a) Define the kind of speaking they want to test; (b) Develop tasks and criteria that test this; (c) Inform the examinees about what they test; and (d) Make sure that the testing and rating processes follow the stated plans (Luoma, 2004).

2.2. Interview

When it comes to assess the oral ability of language learners, the first thing that comes to mind is an oral interview (Brown, 2004). Nevertheless, it has its own merits and demerits. The advocates of the oral interview claim that the examination at least seems to offer a realistic means of assessing the total oral skill in a natural speech context (Heaton, 1989). The interview procedure is not an elicitation technique, but rather a type of framework for applying different elicitation techniques (Madsen, 1983). It can be concluded that probably the most typical format of testing oral interaction is the interview.

Brown (2004) proposed a four-stage framework for oral proficiency interviews including warm-up, level-check, probe, and wind-down. At the warm-up stage, the interviewees are put at ease with common greetings and some easy questions. Then, at the level-check stage, specific questions are put to the interviewees to determine their current level of oral proficiency. At the probe stage, interviewees are thoroughly and meticulously assessed. And finally, at the wind-down stage, the interview is concluded and wrapped up.

According to Weir (1990), the interview procedure can be subdivided into two types including controlled and free. On the one hand, in the controlled interview procedure, the questions are set in advance, more restrictions are imposed and less freedom is given to the interviewees to express their opinions freely because they lack the linguistic competence to do so. On the other hand, in the free interview, the questions are less fixed and more open-ended. Fewer restrictions are imposed on the interviewees and more freedom is given to them to allow them to express themselves more openly.

The interview is the most common of all oral tests and many view it as the only type of oral test (Huxham et al., 2012). Elsewhere, Underhill (1987) divides the interview procedure into two types including a short interview and a long interview. The short interview consists of an introduction phase, a find level phase, and a check questions phase. The long interview consists of the following stages: An introduction and warm-up stage, an establishing approximate level stage, a fine-tuning stage, an eliciting learner's opinion stage, and feedback (invite any comment) and wind-up (end the interview) stage.

There exists abundant literature on the difficulty and significance of assessing oral ability. An absolute majority of the prominent figures in the testing and assessment field have unanimously pointed to the difficulty of assessing speaking and contend that speaking is the most difficult skill to assess reliably (Harris, 1987; Heaton, 1989; Luoma, 2004; Madsen, 1983). Overall, assessing speaking can be regarded as the most challenging of all language skills to prepare, administer and score since no language skill is so demanding to assess with exactness as speaking ability. The problems lie in the discrepancies and inconclusiveness inherent in what criteria to include assessing oral ability, setting tasks that constitute a representative sample of the population of oral tasks. Eliciting a kind of response that genuinely represents the learners' ability, and scoring the samples validly and reliably scored (Brown, 2010: Luoma, 2004, Shehadeh, 2012). To fill the above-mentioned gaps in the literature, this study aimed to investigate the most challenging items (fluency, accuracy, range, interaction, coherence, and pronunciation) from both formative and summative perspectives for EFL learners. Moreover, the formative assessment of the first-rater was compared and contrasted with the summative assessment of the second-rater. Therefore, to achieve the-above-stated objectives, the researchers formulated the following research questions:

1. From a formative perspective, which item (fluency, accuracy, range, interaction, coherence, and pronunciation) posed the greatest challenge to EFL learners?
2. From a summative perspective, which item (fluency, accuracy, range, interaction, coherence, and pronunciation) posed the greatest challenge to EFL learners?
3. Is there any consistency between the formative assessment of the first-rater and the summative assessment of the second-rater?

3. Methodology

3.1. Design and Context of the Study

The present study benefits from a quasi-experimental design and due to administrative difficulties of randomization, convenience or available sampling was utilized and available speaking classes at the University of Qom were taken advantage of.

3.2. Participants

The participants of this study were 46 undergraduate students, 22 males, and 24 females majoring in English Language Literature at the University of Qom. All of the participants were first-year students majoring in English language and literature. Half of the subjects were selected from males and the other half from females to control the impacts of the subjects' gender on the results of the study. Participants met twice a week and for three months.

3.3. Instruments

Two books were used as the required texts for classroom activities. The first one is entitled" *For and against*" authored by Alexander (1968) and the second book written by McCarthy and O'Dell (2013) is entitled "*English vocabulary in use*". Newspaper articles were exploited as well for some classroom activities.

Because the interview procedure is the main technique employed by the researchers of the present study, a separate section is devoted to shedding more light on this important technique of assessing speaking. The scored interview is unquestionably the most commonly applied, and the one with the longest history. Paper-and-pencil tests of pronunciation have been utilized repeatedly for some years, generally jointly with other formats of assessment. The simplest and most commonly used method of measuring speaking is to have one or more trained raters interview each candidate separately and record their overall oral performance. Two semi-structured interviews were conducted as pedagogic tools to serve the purposes of summative and formative assessments done by the two different raters.

3.4. Data Collection Procedure

The following methodological steps were taken to achieve the already-stated objectives of the study. First, some speaking pedagogic tasks (lecturing, discussion, summary-telling, oral quizzes, and newspaper article presentation) were designed and implemented by the researchers. Students were to stick to the tasks and requirements of the class throughout the term as stipulated. Their three-month performances were assessed based on predetermined criteria. The researchers adapted Luoma's (2004) assessment framework including six criteria; fluency, range, coherence, interaction, accuracy, and pronunciation.

Different speaking tasks were included in the research project over a three-month period. Lecturing in the class was one of the speaking tasks to be fully complied with. EFL learners were provided with a model of how to present a lecture in the class at the outset of the speaking class. Participants were given the chance to choose a topic of their interest and present it in the class in front of their classmates when their topic had already been confirmed by the instructor. Prior to lecturing, participants were all told how and based on what criteria they would be assessed. At the end of the lecture, the lecturer was provided with comments from both the instructor and his or her classmates. Throughout the term, each participant was given the lecturing opportunity twice for at most fifteen minutes. To rate the participants' performances on lecturing, fluency, accuracy, range, interaction, pronunciation, and coherence were taken into consideration.

The participants were asked to thoroughly study different units of *for and against* which is written by Alexander (1968) and consists of topics of controversy. The learners were then provided with a model by the teacher of how to present a summary of the unit orally in the class. Following the instructor's model, the learners were required to present their well-prepared summaries. The instructor would go to great lengths to involve all EFL learners in the controversial topics. The EFL learners' summary-telling ability and discussion abilities were assessed throughout the term. The instructor kept a profile of students' progress as well so that the participants would be informed of their weaknesses and strengths whenever necessary.

Each participant was taken two oral quizzes during the semester. They were asked to fully cover two units of the book entitled "*English vocabulary in use"* and based on an already-provided instructor model, present a story of the most applicable vocabularies and collocations. The learners' performances once again were assessed based on the previously-mentioned criteria. The instructor kept a profile of the learners' performances on the quizzes as well.

All the participants were required to present newspaper articles in the class as well. To see how reliable the assessments of the teacher for different speaking tasks (lecturing, discussion involvement, summary-telling, oral quizzes, and newspaper article presentations) throughout the three-month period were, another semi-structured interview was conducted at the end of the semester by a second-rater based on the very same criteria used by the first-rater. However, the first-rater was only engaged in formative assessment and no interview was conducted for the formative rater at the end of the course. The formative assessment of the first-rater was compared and contrasted with the summative assessment of the second-rater to see how much consistency existed between the two. The effectiveness of both summative and formative assessment was investigated by comparing the results gleaned for both with the scores obtained from the interview carried out at the outset of the study prior to the commencement of the speaking course.

A checklist was adapted from Luoma's (2004) framework of assessing speaking according to which the raters were able to make more valid, reliable, and consistent assessments. The researchers were not interested in a psychometric model of assessment which is more limiting, rigorous, and scientific. Rather, a more dynamic and flexible approach to assessment was adopted which more probably resembled performance or alternative assessment. The checklist was taken from analytic descriptors of spoken language (Council of Europe, 2001, pp.28-29) which was cited in Luoma (2004, pp.72-74). The items on the list were accuracy, fluency, range, coherence, pronunciation, and interaction. Level descriptors (A+, A, B+, B, C+, and C) were specified for each item on the checklist according to which the raters assigned scores to the interviewees.

Table 1.

Criteria to Assess Speaking

Accuracy	Fluency	Range	Coherence	Interaction	Pronunciation
A+	A+	A+	A+	A+	A+
A	A	A	A	A	A
B+	B+	B+	B+	B+	B+
B	B	B	B	B	B
C+	C+	C+	C+	C+	C+
C	C	C	C	C	C

Table 2.

Score Range for Each Descriptor

A+	A	B+	B	C+	C
16-20	14-16	11-15	8-11	4-8	0-4

Each participant was given a score based on the previously mentioned descriptions and the level descriptors specified on the checklist. Each item was scored out of twenty.

3.5. Data Analysis Procedure

Luoma’s (2004) rating scale was used for the analysis of oral performances. The rating scale consisted of items like fluency, accuracy, range, interaction, pronunciation, and coherence. Two different raters assessed the participants’ speaking performance based on the above-mentioned criteria: One formatively, during a three-month period based on a classroom schedule that revolved around tasks, and the other summatively, at the end of the semester through an interview based on the same items on the rating scale exploited by the first-rater. MANOVA and pair-wise comparisons were utilized to analyze the data.

4. Results

The data were analyzed through multivariate ANOVA which, besides its specific assumptions, assumes normality of the data. As displayed in Table 3, the ratios of skewness and kurtosis over their standard errors were lower than +/- 1.96, hence normality of the data is assured.

Table 3.

Testing Normality Assumption

		Skewness			Kurtosis	
	Statistic	Std. Error	Ratio	Statistic	Std. Error	Ratio

Formative	Fluency	-.08	.35	-0.23	-.06	.68	-0.09
	Accuracy	-.20	.35	-0.60	-.22	.68	-0.33
	Range	-.36	.35	-1.05	.61	.68	0.89
	Interaction	-.41	.35	-1.19	.78	.68	1.13
	Coherence	-.05	.35	-0.16	1.21	.68	1.76
	Pronunciation	.20	.35	0.57	.73	.68	1.07
Summative	Fluency	-.03	.35	-0.10	.50	.68	0.73
	Accuracy	-.25	.35	-0.74	1.17	.68	1.71
	Range	-.14	.35	-0.42	.69	.68	1.01
	Interaction	-.26	.35	-0.75	.35	.68	0.52
	Coherence	-.22	.35	-0.63	.52	.68	0.77
	Pronunciation	.22	.35	0.63	.78	.68	1.14

The first research question probed into the most challenging item (fluency, accuracy, range, interaction, coherence, and pronunciation) for the participants when they were formatively assessed by the first-rater. The multivariate ANOVA (MANOVA) was run to compare the mean scores of the subjects on the items from a formative perspective. Based on the results displayed in Table 4, it can be concluded that the participants achieved the highest mean on the pronunciation (M = 13.88). This was followed by interaction (13.02), fluency (M = 12.73), accuracy (M = 12.60), range (M = 11.39) and coherence (M = 11.39) respectively.

Table 4.

Descriptive Statistics from a Formative Perspective

Scores	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Fluency	12.73	.45	11.82	13.65
Accuracy	12.60	.18	12.22	12.98
Range	11.45	.42	10.59	12.32
Interaction	13.02	.51	11.98	14.05
Coherence	11.39	.47	10.43	12.34
Pronunciation	13.88	.34	13.20	14.57

The results of multivariate tests F (5, 41) = 22.89, p = .000, Partial η^2 = .736 representing a large effect size) (Table 5) indicated that there were significant differences among the items formatively.

Table 5.

Multivariate Tests from a Formative Perspective

	Effect	Value	F	Hypothesis Df	Error Df	Sig.	Partial Eta Squared
Scores	Pillai's Trace	.736	22.89	5	41	.000	.736

	Wilks' Lambda	.264	22.89	5	41	.000	.736
	Hotelling's Trace	2.79	22.89	5	41	.000	.736
	Roy's Largest Root	2.79	22.89	5	41	.000	.736

The results of pair-wise comparisons (Table 6) indicated that the participants' mean on pronunciation (M = 13.88) and fluency (M = 12.73) (MD = 1.14, p = .031) differed significantly from one another. There was a significant difference between participants' mean on pronunciation (M = 13.88) and accuracy (M = 12.60) (MD = 1.28, p = .003) as well.

Table 6.

Pairwise Comparisons from a Formative Perspective

(I) scores	(J) scores	Mean Difference (I-J)	Std. Error	Sig.^{b}	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Pronunciation	Interaction	.86	.41	.664	-.43	2.16
	Fluency	1.14*	.35	.031	.05	2.23
	Accuracy	1.28*	.31	.003	.29	2.27
	Range	2.43*	.34	.000	1.36	3.50
	Coherence	2.49*	.38	.000	1.31	3.67
Interaction	Fluency	.28	.26	1.000	-.53	1.09
	Accuracy	.41	.48	1.000	-1.07	1.91
	Range	1.56*	.28	.000	.66	2.46
	Coherence	1.63*	.23	.000	.89	2.36
Fluency	Accuracy	.13	.43	1.000	-1.19	1.46
	Range	1.28*	.18	.000	.72	1.84
	Coherence	1.34*	.18	.000	.77	1.92
Accuracy	Range	1.14	.39	.089	-.08	2.38
	Coherence	1.21	.44	.129	-.15	2.58
Range	Coherence	.065	.193	1.000	-.53	.66

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

The participants' mean on pronunciation (M = 13.88), range (M = 11.45) (MD = 2.43, p = .000), and coherence (M = 11.39) (MD = 2.49, p = .000) differed significantly. In the same vein, the participants had a significantly higher mean on pronunciation in comparison with other items. The participants' mean on interaction (M = 13.02) and range (M = 11.45) (MD = 1.56, p = .000) differed significantly as well. From a formative vantage point, the participants' mean on interaction was significantly higher than their mean on other items. The participants' mean on interaction (M = 13.02), coherence (M = 11.39) (MD = 1.63, p = .000), fluency (M = 12.73) and range (M = 11.45) (MD = 1.28, p = .000) differed significantly from one another. Likewise, the participants had a significantly higher mean on fluency compared to their mean on other items of the rating scale, and learners' mean on fluency (M = 12.73) and coherence (M = 11.39) (MD = 1.34, p = .000) differed significantly. The differences among the other pairs of means were insignificant.

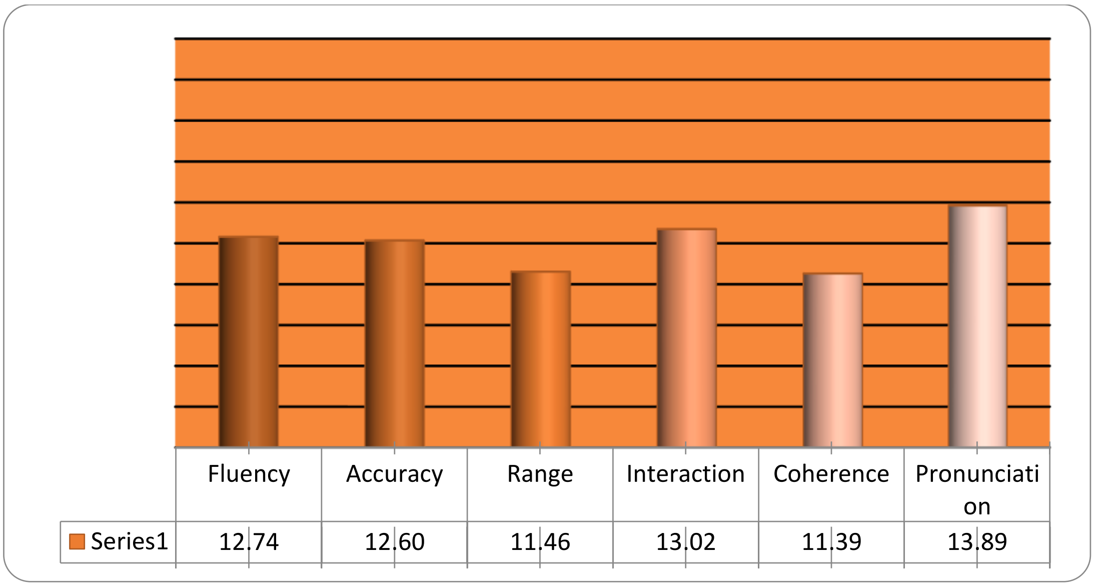

Figure 1. *Means from a Formative Perspective*

The second research question intended to investigate the item (fluency, accuracy, range, interaction, coherence, and pronunciation) which posed the greatest challenge to the participants when they were assessed in a summative way by the second-rater. The multivariate ANOVA (MANOVA) was run to compare the mean scores of the participants on the items from a summative perspective. Based on the results displayed in Table 7, it can be concluded that the learners had the highest mean on the pronunciation item (M = 13.84). This was followed by interaction (12.84), fluency (M = 12.82), accuracy (M = 12.78), range (M = 11.37) and coherence (M = 11.13), respectively.

Table 7.

Descriptive Statistics from a Summative Perspective

Scores	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Pronunciation	13.84	.37	13.08	14.60
Interaction	12.84	.51	11.80	13.89
Fluency	12.82	.42	11.97	13.67
Accuracy	12.78	.40	11.96	13.60
Range	11.37	.44	10.47	12.26
Coherence	11.13	.42	10.26	11.99

The results of multivariate tests (F (5, 41) = 29.18, p = .000, Partial η^2 = .781 representing a large effect size) (Table 8) indicated that there were significant differences among the items from a summative perspective.

Table 8.

Multivariate Tests from a Summative Perspective

Effect		Value	F	Hypothesis Df	Error Df	Sig.	Partial Eta Squared
Scores	Pillai's Trace	.78	29.18	5	41	.000	.781
	Wilks' Lambda	.21	29.18	5	41	.000	.781
	Hotelling's Trace	3.55	29.18	5	41	.000	.781
	Roy's Largest Root	3.55	29.18	5	41	.000	.781

To shed light on differences, pair-wise comparisons (Table 9) were conducted. The results indicated that the participants' mean on pronunciation (M = 13.84) and fluency (M = 12.82) (MD = 1.01, p = .021) differed significantly from one another. In other words, the participants had a significantly higher mean on pronunciation.

Table 9.

Pairwise Comparisons from a Summative Perspective

(I) scores	(J) scores	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Pronunciation	Interaction	.99	.35	.105	-.09	2.08
	Fluency	1.01*	.29	.021	.09	1.94
	Accuracy	1.06*	.33	.038	.03	2.09
	Range	2.47*	.31	.000	1.49	3.46
	Coherence	1.	.30	.000	1.78	3.64
Interaction	Fluency	.02	.27	1.000	-.81	.85
	Accuracy	.06	.31	1.000	-.90	1.03
	Range	1.47*	.28	.000	.59	2.36
	Coherence	1.71*	.25	.000	.94	2.49
Fluency	Accuracy	.04	.17	1.000	-.50	.58
	Range	1.45*	.20	.000	.81	2.09
	Coherence	1.69*	.19	.000	1.08	2.31

Accuracy	Range	1.41^{*}	.19	.000	.80	2.01
	Coherence	1.65^{*}	.23	.000	.93	2.37
Range	Coherence	.23	.19	1.000	-.35	.83

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

The participants' mean on pronunciation (M = 13.84), accuracy (M = 12.78) (MD = 1.06, p = .038), range (M = 11.37) (MD = 2.47, p = .000), and coherence (M = 11.13) (MD = 2.71, p = .000) differed significantly from one another. The difference between participants' mean on interaction (M = 12.84) and accuracy (M = 12.78) (MD = .065, p = 1) was insignificant within the same perspective. The participants had a significantly higher mean on interaction. The difference between their mean on interaction (M = 12.84), range (M = 11.37) (MD = 1.47, p = .000), and coherence (M = 11.13) (MD = 1.71, p = .000) proved to be significant.

Further analysis revealed that the learners' mean on fluency turned out to be significantly higher than their mean on other items. The difference between participants' mean on fluency (M = 12.82), range (M = 11.37) (MD = 1.45, p = .000), and coherence (M = 11.13) (MD = 1.69, p = .000) was significant. The difference between participants' mean on accuracy (M = 12.78), range (M = 11.37) (MD = 1.41, p = .000), and coherence (M = 11.13) (MD = 11. 65, p = .000) proved to be significant as well. The learners had a significantly higher mean on accuracy too. The differences among the other pairs were insignificant.

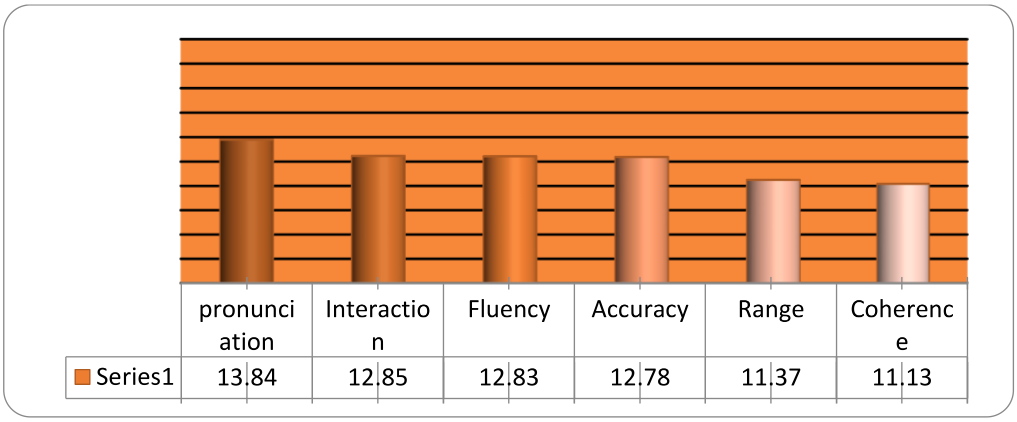

Figure 2. *Means from a Summative*

The third research question probed into the consistency between the formative assessment of the first-rater and the summative assessment of the second-rater. In other words, the researchers intended to know whether there was an agreement between the two forms of assessment when the same rating scale was adopted by the two raters.

The Pearson correlation was run to probe into the consistency between the formative assessment of the first-rater and the summative assessment of the second-rater. Based on the results displayed in Table 10, it can be concluded that there was a significant agreement between the summative and formative rating of participants in terms of their pronunciation (r (44) = .84, p = .000, representing a large effect size), interaction (r (44) = .95, p = .000, representing a large effect size), fluency (r (44) = .93, p = .000, representing a large effect size), range (r (44) = .93, p = .000, representing a large effect size), and coherence (r (44) = .94, p = .000, representing a large effect size).

Table 10.

Pearson Correlation for the Consistency between Formative and Summative Assessment

		SumPro	SumInter	SumFlu	SumAccu	SumRange	SumCoh
FormPro	Pearson Correlation	.848**					
	Sig. (2-tailed)	.000					
	N	46					
FormInter	Pearson Correlation		.951**				
	Sig. (2-tailed)		.000				
	N		46				
FormFlu	Pearson Correlation			.930**			
	Sig. (2-tailed)			.000			
	N			46			
FormAccu	Pearson Correlation				.335*		
	Sig. (2-tailed)				.023		
	N				46		
FormRange	Pearson Correlation					.930**	
	Sig. (2-tailed)					.000	
	N					46	
FormCoh	Pearson Correlation						.945**
	Sig. (2-tailed)						.000
	N						46

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

The consistency between the summative and formative rating of participants with regard to fluency (r (44) = .33, p = .023, representing a moderate effect size) proved to be significant but moderate.

5. Discussion

The present study was an attempt to compare and contrast the oral production of a group of EFL learners through formative assessment of the first-rater with the summative assessment of the second-rater based on the same rating scale. The researchers also aimed to ascertain the most challenging item of the rating scale (fluency, accuracy, range, coherence, interaction, and pronunciation) for language learners from both summative and formative perspectives. Therefore, the study set to find out whether a consistency exists between the two forms of assessment carried out by two different raters based on the same criteria.

This study adapted Luoma's (2004) speaking model as a theoretical framework and as a point of reference. The overall findings of this study, consistent with a compelling body of evidence from previous studies (e.g., Black & William, 1998; Clarke, 1998; Jones, 2005; Norris, 2016; Sadler, 1998; Torrance & Pryor, 1998, Weaver, 2012; Wigglesworth & Frost, 2017) support the position that the application of both forms of assessment (formative and summative) could be facilitative for second or foreign language teachers and learners and there exists a consistency between two forms of language assessment (summative and formative)

Tuan (2012) sought to discern if the analytic scoring approach would be more instrumental in enhancing learners’ speaking performances. The study ascertained that the students held a positive attitude towards the adoption of the analytic scoring approach in teaching and assessing speaking skills which are consistent with the application of both summative and formative assessment based on some predetermined criteria.

The results of the present study are in line with the findings of many researchers who have advocated the utilization of both types of assessment in second or foreign language classroom contexts (e.g., Black & William, 1998; Black et al., 2003; Clarke, 1998; Jones, 2005; Sadler, 1998; Torrance & Pryor, 1998). Regarding the pedagogic and target speaking tasks were utilized in this study, the results revealed that manipulating pedagogic and target tasks by using instructional materials and teacher modeling can be instrumental in promoting learners' awareness of their overall progress in speaking, and this finding ratifies the previous research done on learners' oral proficiency level employing authentic, real-world tasks (Brindley, 2013; Bygate, 2016; Knight, 1992; Long, 2015; Norris, 2016). The learners were provided with a reporting card throughout the semester as the formative assessment was carried out so that they were informed of the amount of progress they had made in speaking. At the end of the semester, when they were summatively assessed by a second-rater, a real comparison could be made by all the participants about their overall progress in speaking about items like fluency, accuracy, range, pronunciation, and coherence.

The present study revealed that there existed a consistency between the two forms of assessment (formative and summative) though the raters were different. One possible explanation for the above-mentioned agreement can be attributed to the fact the same items (fluency, accuracy, range, interaction, pronunciation, and coherence) were taken into account when the raters had to carry out the assessments. The raters had to comply with the same specifications for each item on the rating scale in formative and summative assessment alike.

Another possible justification for the consistency between the two types of assessment can be because all language learners were already familiarized with the items on the rating scale prior to the commencement of the course. This familiarity might have served as a crucial point of reference for language learners. This in turn might have led to this agreement between formative and summative assessment.

From both summative and formative perspectives, pronunciation was the least challenging item. This could be because most foreign language learners are interested to attain native-like pronunciation and this passion to attain native-like pronunciation helps them become intrinsically motivated. Moreover, the technological boom and the availability of educational materials on the Internet, mobiles, TVs, and language institutes have been quite instrumental in exposing language learners to appropriate target language pronunciation. This amount of exposure might have raised their interest in adopting a positive attitude towards the native-like mastery of pronunciation as well.

From both summative and formative assessment perspectives, coherence and range posed the greatest challenge to foreign language learners. This can be attributed to the fact that language learners need to master a wide range of target language structures, to have great flexibility in reformulating ideas with various linguistic forms to convey meaning, emphasize, differentiate, and remove ambiguity, and have a good command of idiomatic expressions and colloquialisms as well to possess great range in one’s speech (Luoma, 2004). These might explain why range posed a major challenge to language learners. From both summative and formative assessment vantage points, coherence was the most challenging item for language speakers to attain because they had to create coherent links and cohesive discourse markers within the speech patterns; they also had to apply a variety of organizational patterns and a wide range of connectors and other cohesive devices (Luoma, 2004).

6. Conclusion

Assessment when properly apprehended and not mistaken for other testing terminologies such as measurement and tests can be viewed as a driving and challenging force for language learners within the classroom environment. Much of what transpires in EFL classes can be regarded as assessment since students are primarily involved in pair work, group work, classroom discussions, quizzes which in one way or another help them unleash their full potential in speaking. Formative and summative assessments form an essential part of the teaching and learning process. They are being applied to provide the pupils with the necessary feedback to promote learning and help the teacher understand students' learning. They can provide a vivid picture of students' progress along the way.

The principled and systematic application of both types of assessment (summative and formative) is consistent with current second language methodologies such as task-based language teaching (Brown, 2010; Gipps, 1994). Drawing on Gipps' educational model of language assessment (1994), educators must move beyond what he describes as a psychometric model of language testing and strive for a more dynamic model of language assessment that is performance-based and process-oriented. Therefore, these two types of assessment can provide language teachers with insightful information that can act as helpful diagnostic tools to help language learners remedy their language-related problems. Accordingly, language learners will be informed of their potential weaknesses and strengths as a result of being assessed in a summative and formative way.

This research provides an insight into the consistency and interconnection between the formative and summative assessment and students' improvement in speaking learning. This insight prompts teachers to adopt a clear-cut rating scale to assess their students' speaking performance. Moreover, this research indicated that assessing students' speaking performance should be viewed as a process rather than a product. This research is a reaction to the sole application of the holistic scoring approach in teaching and assessing speaking skills. The holistic scoring approach can bring some benefits to teachers in teaching and assessing students’ speaking performance but still questions the autonomy and continuity of the learning speaking as a process.

A survey of the pedagogical assessment in general and speaking assessment in particular in the field of language teaching demonstrates that a substantial part of these studies have been concerned with formative and summative approaches, mainly generalizing that both language learners and teachers can benefit from the application of theses two formats of speaking assessments in educational settings. The results of this study revealed that the employment of formative and summative speaking assessment was beneficial for language learners in terms of helping them along the way of improving their overall speaking performances, and instrumental in finding the needed remedial materials for language teachers and practitioners. The opportunities that these two formats of assessment can provide for the ESL or EFL instructors and pedagogical interventionists can be regarded as a rich and underexplored area that needs more attention. Today, many researchers and practitioners advocate the principled and judicious application of both formative and summative assessments in language classrooms. The results of this study imply that by taking optimal advantage of both summative and formative types of assessment, particularly in EFL settings, both learners and teachers can be guided appropriately. Therefore, it is hoped that the application of both summative and formative assessment will rectify, enhance and enrich the status quo within the language testing and language teaching domain.

The study suffers from some limitations. First and foremost, due to administrative difficulties, only a small sample size was included which undermines the generalizability of the findings. Larger sample size can be included to boost the generalizability of the findings to other contexts. The second limitation of the study refers to the instrumentation and scoring procedures adopted. Techniques other than the interview can be used and frameworks other than that of Luoma (2004) can be adopted as a point of reference to assess speaking. Third, this study was done in an EFL context which might fail to be generalized to an ESL context. Future studies can be conducted in ESL settings to find out whether the same results are achievable. More specifically, studies can be carried out with a more specific emphasis and focus on subcategories and subcomponents of speaking ability. Some new innovative scaffolding techniques of improving the overall speaking ability of EFL and ESL language learners can be utilized to discern the amount of progress that can be made with and without recourse to them. Moreover, comparison and contrast of summative and formative assessment in other language skills such as listening and writing can be an interesting venue of research for interested researchers within the field of applied linguistics to further delve into.

References

Alexander, L. G. (1968). *For and against.* Longman Publications.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. & Palmer, A. (1996). *Language testing in practice*: *Designing and developing useful language tests*. Oxford University Press.

Black, P. J., & Wiliam, D. (1998). *Inside the black box. Raising standards through classroom assessment.* King's College.

Brindley, G. (2013). Task-based assessment. In. C. Chapelle (Ed.), *The encyclopedia of applied linguistics*, (pp. 1-6). John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal1141

Bygate, M. (2016). Sources, developments, and directions of task-based language teaching. *The Language Learning Journal*, *44*(4), 381-400. https://doi.org/10.1080/09571736.2015.1039566

Clarke, S (1998). *Targeting assessment in the primary school*. Hodder and Stoughton.

Brown, H. D. (2010). *Language assessment: Principles and classroom practices*. Longman.

Brown, J. D. (2004). *Understanding research in second language teaching: A teacher's guide to statistics and research design*. Cambridge University Press.

Chapelle, C. A., & Brindley, G. *(*2010*). Assessment. In* N. Schmitt *(*Ed.*),* An introduction to *applied linguistics (*pp. 247-267*).* Hodder Education.

Gipps, V. C. (1994). *Beyond testing*. Farmer Press.

Harris, P. D. (1987). *Testing English as a second language*. Hill Book Company.

Heaton, J. B. (1989). *Writing English language tests*. Longman.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Journal of Assessment and Evaluation in Higher Education*, *37*(1), 125-136.

Jafarpur, A. (1994). *A Course in language testing*: Payame-Noor University Press.

Jones, J. (2005). Developing effective formative assessment practices in the primary modern foreign language (MFL) classroom. *Encuentro, 15*, 39-47.

Knight, B. (1992). Assessing speaking skills: A workshop for teacher development. *ELT Journal, 46*(3), 294-302.

Long, M. (2015). *Second language acquisition and task-based language teaching*. Wiley- Blackwell.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Madsen, S. H. (1983). *Techniques in testing*. Oxford University Press.

McCarthy, M. & O'Dell, F. (2013). *English vocabulary in use.* Cambridge University Press.

McNamara, T. & Hill, K. (2011). Developing a comprehensive, empirically-based research framework for classroom-based assessment. *Language Testing*, *29*(3), 395-420. https://doi.org/10.1177/0265532211428317

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230-244.

Richards, J. C. & Renandya, W. A. (*2002*). *Methodology in language teaching*: *An anthology of current practice.* Cambridge University Press.

Sadler, D. R. (1998). Formative assessment: revisiting the territory. *Assessment in Education, 5*(1), 77-84.

Shehadeh, A. (2012). Task-based language assessment: Components, development, and implementation. *The Cambridge Guide to Second Language Assessment, 4,* 156-163.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment. Teaching, learning, and assessment in the classroom.* Open University Press.

Tuan, L. (2012). Teaching and assessing speaking performance through the analytic scoring approach. *Theory and Practice in Language Studies*, *2*(4), 673-679.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge University Press

Weaver, C. (2012). Incorporating a formative assessment cycle into task-based language teaching in a university setting in Japan. In A. Shehadeh & C. Coombe (Eds.), Task-based language teaching in foreign language contexts: Research and implementation, 287-309. John Benjamins. https://doi.org/10.1075/tblt.4.17wea

Wigglesworth, G. & Frost, K. (2017). Task and performance-based assessment. In E. Shohamy, S. May, & I. Or. (Eds.), Language testing and assessment: Third edition. Encyclopedia of Language and Education (pp. 121-130). Springer. https://doi.org/10.1007/978-3-319-02261-1_8

Weir, J. C. (1990). *Communicative language testing*. Prentice Hall