Research in English Language Pedagogy

RELP (2019) 7(2): 187-208 DOI: 10.30486/relp.2019.665887

The Impact of Task Types and Rating Methods on Iranian EFL Learners' Speaking Scores

Soghra Ahangari

Department of English Language Teaching, Islamic Azad University, West Tehran Branch, Tehran-Iran

Email address: Sara.ahangari93@gmail.com

*Shokouh Rashvand Semiyari**

Department of English Language Teaching, Islamic Azad University, East Tehran Branch, Tehran-Iran

Email address: Sh.Rashvand@iauet.ac.ir

Abstract

Speaking assessment is quite challenging as there are many factors in addition to speakers' ability contributing to how well someone can speak a language. In this study, 40 male and female upper intermediate EFL learners, with age range of 15-26, selected through ECCE test. They performed four different types of speaking tasks (explaining, problem-solving, story-telling, and picture-describing). These tasks were rated by two raters using two scoring methods: holistic and analytic. The one-way repeated measures ANOVAs, paired-sample t-tests and Pearson Product Moment Correlations illustrated that the task types and rating methods didn't have any significant effects on learners' speaking scores as far as problem-solving, explaining, and picture-describing tasks were concerned. Yet, the rating methods represented to have some effects on story-telling task. The findings also indicated a significant correlation between holistic and analytic ratings of the problem-solving, picture-describing and story-telling tasks while the reverse was true for the explaining task.

Keywords: assessment, holistic and analytic rating methods, speaking scores, task types.

* Corresponding Author Submission date: 3 April, 2018 Acceptance date: 13 Oct, 2018

1. Introduction

With the growing popularity of Communicative Language Teaching (CLT), the major goal of the most L2 programs is to develop speaking ability. As Brindley (1998) notes, assessing speaking is challenging because our impressions of students' speaking assessment are affected by many factors. Since speaking skill has a decisive role in language teaching curriculum, its assessment has become an important aspect of many studies in recent years. Prabhu (1987) proposed that engaging students in language tasks will certainly lead to more effective learning as they may go beyond the pure language. Task-based teaching is making chances for promoting real meaning-focused language use and increasing learners' participation (Ellis, 2003). Tasks are currently investigated for their relative impacts on learning and their educational commitments to classroom and out-of-class acquisition (Luoma, 2004). As proposed by Luoma (2004), connecting rating scales to peculiar tasks is the unavoidable end result. He also emphasized that in addition to all the agents that may influence learners' performance, the scale used to evaluate the performance can diversify extremely from global assessments to detailed analytic scales. In fact, the way in which these scales are described by an assessor may have an impact on the score or scores awarded to the test takers. Despite all these studies, there is no study that dealt with the implementation of speaking assessment in an Iranian EFL context. Though some studies have been done regarding the impacts of some task types such as picture-description and story-presentation on speaking assessment (Teng, 2007), there are few -if none- studies regarding the effects of problem-solving, story-telling, picture-describing and explaining tasks on learners' speaking performance. This study, therefore, examined the impacts of task types and rating methods on learners' speaking scores. To this end, the current study was intended to answer the following questions:

1. Do task types have any effects on Iranian EFL learners' speaking scores?
2. Do rating methods have any effects on Iranian EFL learners' speaking scores?
3. Is there any relationship between rating methods and task types?

2. Literature Review

2.1. Tasks

Task is a tool whereby learners participate in language use. It emphasizes the meaning and accordingly paves the way for making the opportunities for language

acquisition (Ellis, 2003). According to Bygate, Skehan, and Swain (2001), a task is an activity for learners to use meaning–focused language to achieve an objective.

Ellis (2001) and Swain (1995) pointed out that the language structure has been focused greatly in second language (L2) pedagogy and L2 research for more than 20 years. For this reason, tasks or task-based activities have been frequently dealt with in SLA journal articles, book titles, and conferences. As Bygate (1999) put forward, it may cause sound principles for designing classroom materials empirically. Teachers can choose task designs and performance conditions intentionally to draw learners' attention to special parts of language being learned. Knowing how to do a special kind of task will direct learners' performance in expected ways. It may also develop learning opportunities and promote their proficiency (Candlin, 1987; Skehan, 1998 and Samuda, 2001).

Learning-centered education that supports the meaning-focused activities consists what Prabhu (1987) announced: "(a) information-gap, (b) reasoning-gap, and (c) opinion-gap tasks". Richard and Rodgers (2014) describe *information-gap* activities include the exchange of information among learners to carry out a task. They further explain an *opinion-gap exercise* needs learners to discuss and exchange their particular choices, emotions, or opinions to carry out a task. They do not require to reach agreement. Prabhu (1987) describes a *reasoning-gap* activity needs learners to run some new information by deducing it from information they have been provided.

Long (2015) believed that task-based language teaching deals with the way people learn languages and the social values. He further added that it underlies the basis in philosophy of education and meets the requirements for accountability, relevance, avoidance of known problems with existing approaches, learners-centeredness, and functionality. Ellis (2006) highlighted three approaches in tasks design including task design in Direct System Referenced -Tests, Design in Direct Performance-Reference Tests, and Integrating the two approaches to Task Design.

Jackson (2011) studied the different learner language production via convergent and divergent tasks. Lim and Lee (2015) investigated background and perceptions on task performance (convergent vs. divergent tasks) of Korean participants under different modes (face to face conversations vs. mobile chatting). The results confirmed the effect of both modality and task types on conversational changes.

2.2. Rating Methods

There are two major scoring methods; holistic and analytic scoring. "Analytic scoring is a kind of assessment when the objectives of the final product are broken down into criteria parts, and each part is scored independently" (Tuan, 2012, p.2). Park and Park (2004) points out, "holistic scores give students a single, overall assessment score, using one evaluation item for the assessment of a large number of test takers in a limited amount of time" (p.1). Concerning when to use holistic rubrics, Airasian and Russell (2001) pointed out that here is no particular accurate response to a task (e.g., creative work), the emphasis is usually put on general quality, efficiency, or understanding of a certain content or skill, the evaluation is often summative (e.g., at the end of a semester or major), and you are evaluating significant numbers (e.g., 150 Senior portfolios). In evaluating rubrics, several faculties evaluating the learners' performance to develop a consistent scoring. Outside audiences will be examining rubric scores.

Wiseman (2012) compared the performance of a holistic and an analytic scoring rubric to evaluate ESL writing for placement and diagnostic objectives in a community college fundamental skills program. Providing evidence of reliability and validity for both analytic and holistic rubrics indicates that analytic rubrics are more suitable to separate test takers for distinguishing and placing aims. Ghalib and Al-Hattami (2015) compared the accomplishments of holistic and analytic scoring methods in an EFL writing task situation. The findings determined the reliability and validity of both methods. Analytic scoring rubrics, however, put the test takers along a more obviously determined scale of writing proficiency, and were, therefore, more reliable than holistic scoring rubric devices. Chi (2001) compared holistic and analytic scoring methods to examine how the alternate scorings can make differences for performance assessment using many-faceted Rasch model. He revealed that the selection of scoring methods might not be significant for the relative comparison of students but it could have serious implications for the assessment of students' absolute abilities. For rater consistency, analytic scorings contributed more consistently than holistic scorings.

2.3. Assessment

Assessment is a procedure of assembling data in order to make decisions about individuals and groups (Salvia, Ysseldyke, and Bolt, 2007). Assessment tasks are

considered as necessary tools for eliciting and evaluating task performance of learners that are meaning focused for certain goals (Ellis, 2003). Brindley (1994) proposed the definition of task-centered language assessment which is sufficiently general to cover both in-class and out-of-class conditions and can be used to the evaluation of efficiency needed independently of the curriculum or to curriculum-based performance. Task-centered language assessment is the procedure of assessing, dealing with a set of explicitly asserted principles, the quality of the interactive performance drawn out of students as part of aim-directed, meaning-focused language practice needing the combination of skills and knowledge.

Mc. Grow (as cited in Teach Thought Staff, 2015) listed 6 types of assessment as diagnostic assessment (assesses strengths, weaknesses and knowledge of students before instruction), formative assessment, summative assessment, norm-referenced assessment (compares a student's performance with that of others including national groups or other "norms"), criterion-referenced assessment (evaluates a student's performance against a goal, specific objective, or standard), and interim/benchmark assessment (assessing students' performance at periodic intervals, frequently at the end of a grading period).

Colomar (2014) demonstrated and examined a classroom-based assessment system to investigate students' speaking abilities in different skilled contexts in tourism. To increase validity and reliability, he decided to create an assessment procedure based on a combination of testing formats, rating criteria and rating scales. He examined the effect of self/peer assessment on civil designing instruction. He found that students who carried out self/peer assessment during diverse learning activities obtained better performance.

In a few past years, some researchers have investigated the effects of task types and rating methods on second language speaking assessment (Lee, 2006; Lumley & Sullivan, 2005). Upshur and Turner (1999) studied the systematic effects in the rating of second-language speaking ability. The findings were consistent with LT research into systematic effects of task and rater on ratings and with SLA research into systematic effects of task on discourse. They inferred that task type influences strategies for assessing language performance.

Eckes (2005) examined rater effects in the writing and speaking sections of the Test of German as a Foreign Language (TestDaF). The focus was on rater main effects as well as 2- and 3-way interactions between raters and the examinees, rating criteria (in the

writing section), and tasks (in the speaking section). Results showed that raters differed strongly in the severity with which they rated examinees; they were fairly consistent in their overall ratings; and they were substantially less consistent in relation to rating criteria (or speaking tasks, respectively) than in relation to examinees.

Teng and Huei-Chun (2007) investigated the effect of task type on the performance of EFL speaking tests for Taiwanese college students. They indicated that there was no significant difference in the subjects' holistic rating scores for the three task types, including answering questions, picture description, and presentation. That is, test takers did not perform differently on various task types of EFL speaking test. However, significant main effects were found for task type on the two analytic measures, i.e., complexity and fluency.

In'nami and Koizumi (2015) examined Task and Rater effects in L2 speaking and writing through Generalizability Studies. They found out that the most of scores' variation was due to the examinees' performance. They also showed that task and task-related interaction effects explained a greater percentage of the score variances, than did the rater and rater-related interaction effects.

Han and Huang (2017) studied the impact of scoring methods on the institutional EFL writing assessment. They indicated that holistic scoring can produce as reliable and dependable assessment outcomes as analytic scoring. They also showed that all raters prefer using the holistic scoring method because it could help them not only assign fair and objective scores to essays but also facilitate their scoring process. Moreover, most raters agreed that the content of an essay (i.e., type of tasks) was the most important factor that most affected their holistic scorings. While in contrast, all aspects of an essay (e.g., grammar, content, or organization) jointly affected their analytic scorings.

3. Methodology

3.1. Design and Context of the Study

The design of the study was quasi-experimental. In a quantitative descriptive-analytic study two independent variables contributed to this study; four task types (problem-solving, picture- describing, story-telling, and explaining) and two rating methods (holistic vs. analytic). The dependent variable was Iranian EFL learners' speaking scores. This study was done in English Hut Institute, Mofid and Roshangar high schools in Tehran.

3.2. Participants

This study included 40 male and female upper intermediate participants. The age range of these students was 16-25. They were studying English language in English Hut Institute, Mofid and Roshangar high schools in Tehran. They were selected based on their results on the ECCE Michigan Test. These learners were truly homogenous with regard to their English proficiency levels. This study also enjoyed two sets of raters. One of them was one of the researchers and other ones were the teachers of the afore-mentioned institute/high schools. It should be reminded that the participants' gender and age were not considered as independent variables of the study. Table 1 demonstrates the demographic data of the participants:

Table 1.

Demographic Background of the Participants

No. of Students	40
Gender	20 females & 20 males
Native Language	Persian
Field of Study	EFL
Institute/High Schools	Hut/ Mofid & Roshangar
Academic Year	2016-2017

3.3. Instruments

To obtain the results of the study, the researchers used the subsequent instruments:

3.3.1. Cambridge Michigan ECCE Test

The Examination for the Certificate of Competency in English (ECCE) is a standardized high-intermediate level English as a foreign language (EFL) examination. It is developed and scored by the English Language Institute of the University of Michigan (ELI-UM) and has been administered by over 130 authorized test centers around the world. The content and difficulty of ECCE aim at B2 level (independent user). Participants were selected based on their results on the ECCE Test and they indicated to be homogeneous. After obtaining the results of the proficiency test, only those participants whose scores were one standard deviation above and below the mean were chosen as the sample of the study. The ECCE emphasized communicative use of English rather than a formalistic

knowledge of English and it captured the students who were able to function and perform communicative transactions in all four skill areas of the language (speaking, listening, reading, and writing). The language test was 'general', rather than 'academic'. The test took 3 hours and it had different sections including Listening Comprehension including 50 questions, Grammar comprising 35 questions, Vocabulary bearing 35 questions, Reading Comprehension having 30 questions, Writing possessing 1 task and Speaking involving an interview with the examiners.

3.3.2. Tasks

Table 2.

The Four Speaking Tasks, Task Names, Communicative Acts and Tasks Description

Tasks Names	Communicative Acts	Tasks Description
A	Explaining	To explain to a patient what is going to happen before surgery.
B	Problem-Solving	To solve the problem of traffic in town center. To think of three alternative solutions and decide on the cheapest, the most innovative and the most environmentally friendly ones.
C	Picture-Describing	To give each student a photo or picture and having her/him describe what s/he sees in the picture.
D	Story-Telling	To tell about a special event in their life or tell about a perfect day they have had.

This study enjoyed four speaking tasks including explaining, problem-solving, picture-describing, and story- telling. These tasks are fully described in Table 2.

3.3.3. Holistic and Analytic Ratings

All tasks were scored twice, once holistically, based upon general impression on language use, using the revised scale for the British Council's ELTS test (Hughes, 1989) and once analytically, according to LAAS analytic rating scale focusing on pronunciation,

vocabulary, cohesion, organization, and grammar (Sawaki, 2007). Each participant's speech was carefully analyzed for the five aspects mentioned in the rating scale. Particular sub-skills such as using appropriate intonation, turn-taking, using context specific vocabulary, and rate of speech were also considered in assigning the scores.

3.4. Data Collection Procedure

The homogenous participants were supposed to do four tasks on communicative functions as Table 1 indicates. Twenty minutes were devoted to each task. All tasks were scored twice, once holistically and once analytically. First, the researchers went through *Quantitative Data Collection* in a way the participants' speaking tasks were scored holistically. Second, they applied *Quantitative Analysis* in a sense, the participants speaking tasks were scored with respect to the LAAS analytic rating scale (Sawaki, 2007). Such analysis was both linguistic and statistical. Spoken responses were coded for linguistic features with respect to each of the five dimensions of the rating scale. The current study captured the variation in the responses and created a representative sample by randomly choosing 20% of the whole data as already confirmed by Kim, Payant, and Pearson (2015). There were two sets of raters. One of the raters was one of the researchers of the present study. Second raters were the teachers of English Hut Institute, Mofid, and Roshangar high schools. To examine the inter-rater reliability, %20 of tasks were rated by two sets of raters.

3.5. Data Analysis Procedure

The researchers aimed at investigating whether methods of rating, types of task and their correlation had any significant effects on the performance of the participants on the speaking test. To this end, two steps of data analyses, Repeated-Measures ANOVA, four paired sample T-test were conducted.

4. Results

The analysis results on the participants' tasks and the rating methods revealed what follows: first, two one-way repeated measures ANOVAs were conducted on students' holistic and analytic ratings of the four task types. Before conducting ANOVAs, the normality of the ratings across different tasks and rating methods was investigated and all

the skewness values were between -2 and +2, so the data were normal and suitable for the analysis. The descriptive statistics of analytic ratings of different task types is provided below (See Table 3). It should be mentioned that in all the analytic ratings, the average of rating (5-point Likert scale in which 5 shows very good and 1 indicates very poor performance) of different components (vocabulary, grammar, pronunciation, cohesion, and organization) was computed and used in the different stages of data analysis.

Table 3.

Descriptive Statistics of Analytic Ratings of Different Task Types

Task Types	Mean	Std.	N
A Problem-Solving Task	3.99	0.87	40.00
A Picture-Describing Task	4.15	0.76	40.00
A Story-Telling Task	4.21	0.68	40.00
A Explaining Task	4.31	0.60	40.00

As Table 4 indicates, ANOVA analysis revealed that there were not any significant differences among the analytic ratings of tasks, F (3, 90) = 1.48, p = .23. It can be claimed that the type of tasks did not have any effects on the analytic ratings of students' speaking ability.

Table 4.

Repeated-Measures ANOVA: Tests of Within-Subjects Effects in Analytic Rating

Source		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Analytic	**Sphericity Assumed**	**1.64**	**3.00**	**0.55**	**1.48**	**0.23**	**0.05**
	Greenhouse-Geisser	1.64	2.58	0.63	1.48	0.23	0.05
	Huynh-Feldt	1.64	2.85	0.58	1.48	0.23	0.05
	Lower-bound	1.64	1.00	1.64	1.48	0.23	0.05
Error (Analytic)	Sphericity Assumed	33.23	90.00	0.37			
	Greenhouse-Geisser	33.23	77.50	0.43			
	Huynh-Feldt	33.23	85.44	0.39			
	Lower-bound	33.23	30.00	1.11			

Before conducting the second repeated-measures ANOVA for comparing the holistic ratings across different task types, the normality of the ratings across different tasks and rating methods was investigated and all the skewness values were between -2 and +2. Thus the data were normal and suitable for the analysis. The descriptive statistics of holistic ratings of different task types is provided below (See Table 5).

Table 5.

Descriptive Statistics of Holistic Ratings of Different Tasks

Task Types	Mean	Std.	N
HProblem-Solving	4.20	1.06	40.00
HPicture-Describing	4.37	0.81	40.00
HStory-Telling	4.57	0.57	40.00
HExplaining	4.33	0.80	40.00

Note. A = Analytic, H = Holistic

The analyses indicated that there were not any significant differences among the holistic ratings of tasks, F (3, 87) = 1.15, p = .33 >.05. Therefore, it can be claimed that the type of tasks did not have any effects on the holistic ratings of students' speaking ability either. Tables 6 demonstrates the results.

Table 6

Repeated-Measures of ANOVA: Tests of Within-Subjects Effects in Holistic Rating

Source		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Holistic	**Sphericity Assumed**	**2.067**	**3**	**.689**	**1.154**	**.332**	**.038**
	Greenhouse-Geisser	2.067	2.697	.766	1.154	.330	.038
	Huynh-Feldt	2.067	3.000	.689	1.154	.332	.038
	Lower-bound	2.067	1.000	2.067	1.154	.292	.038
Error (Holistic)	Sphericity Assumed	51.933	87	.597			
	Greenhouse-Geisser	51.933	78.210	.664			
	Huynh-Feldt	51.933	86.997	.597			
	Lower-bound	51.933	29.000	1.791			

Second, four paired sample T-Tests were conducted. Descriptive statistics of different ratings of different task types is given in Table 7. As it can be seen in Tables 7 and 8, there was no significant difference between holistic and analytic ratings of the problem-solving, t (29) = -1.35, p = .19 > 0.05, explaining, t (30) = .29, p = .77 > 0.05, and picture-describing tasks, t (33) = -1.99, p = .06 > 0.05. However, a significant difference was found between holistic and analytic ratings of the story-telling task, t (31) = -3.82, p = .00 < 0.05. The holistic rating was higher (M = 4.41, SD = .84) than the analytic one (M = 4.16, SD = .74). It can be therefore concluded that the rating method did not have any effects on students' speaking ratings in problem-solving, explaining, and picture-describing tasks but it had some effects on students' speaking rating in the story-telling task. Tables 7 and 8 depict the results;

Table 7.

Descriptive Statistics of Different Ratings of Different Task Types

	Task Types	Mean	N	Std. Deviation	Std. Error Mean
Pair 1	AProblem-Solving	4.00	40.00	0.89	0.16
	HProblem-Solving	4.20	40.00	1.06	0.19
Pair 2	APicture-Describing	4.03	40.00	0.87	0.15
	HPicture-Describing	4.24	40.00	0.96	0.16
Pair 3	AStory-Telling	4.16	40.00	0.74	0.13
	HStory-Telling	4.41	40.00	0.84	0.15
Pair 4	AExplaining	4.31	40.00	0.60	0.11
	HExplaining	4.26	40.00	0.89	0.16

Table 8.

Paired Samples t-tests of Tasks across Rating Methods

		Paired Differences						
		Mean		95% Confidence Interval of the Difference				
			Std.	Lower	Upper	t	df	Sig.
Pair 1	AProblem-Solving HProblem-Solving	-0.20	0.81	-0.50	0.10	-1.35	29.00	0.19
Pair 2	APicture-Describing HPicture-Describing	-0.21	0.60	- 0.42	0.00	-1.99	33.00	0.06
Pair 3	AStory-Telling HStory-Telling	-0.25	0.37	-0.38	-0.12	-3.82	31.00	0.00
Pair 4	AExplaining HExplaining	0.05	0.99	-0.31	0.42	0.29	30.00	0.77

Third, four Pearson Product Moment Correlations were utilized. As table 9 indicates, there was a significant correlation between holistic and analytic ratings of the problem-solving, r = .66, p = .00, picture-describing r = .78, p = .00, and story-telling tasks, r = .89, p = .00 (See Table 6). However, there was no significant correlation between holistic and analytic ratings as far as explaining task was concerned, r = .15, p = .39.

Table 9.

Paired Samples Correlations between Analytic and Holistic Ratings in ifferent Tasks

	Task Types	N	Correlation	Sig.
Pair 1	AProblem-Solving HProblem-Solving	40	.667	.000
Pair 2	APicture-Describing HPicture-Describing	40	.785	.000
Pair 3	AStory-Telling HStory-Telling	40	.897	.000
Pair 4	AExplaining HExplaining	40	.157	.399

The inter-rater reliability of two sets of raters in holistic ratings are also represented as Table 10.

Table 10.

The Inter-Rater Reliability of Two Sets of Raters in Holistic Ratings

Task Types	Interrater-Reliability
Explaining	% 95
Problem-Solving	% 82
Story-Telling	% 92
Describing	% 78

As Table 10 illustrates, the inter-rater reliability index for the raters in explaining was estimated as %95, in problem-solving as %82, and in story-telling task as %92. These results suggest that there is an almost perfect/significant agreement between the holistic ratings of two sets of raters for each of aforementioned task. The inter-rater reliability index for the raters in describing-task however was estimated as %78 which shows a substantial agreement. The inter-rater reliability of two sets of raters in analytic ratings are reported as Table 11;

Table 11.

Inter-Rater Reliability of Two Sets of Raters in Analytic Ratings

Task Types	Interrater- Reliability
Explaining	%73
Problem-Solving	%92
Story-Telling	%88
Describing	%76

As it is displayed in Table 11, the inter-rater reliability index for the raters in explaining-task was estimated as % 73, which shows a substantial agreement, in problem-solving as %92 and in story- telling task as %88, which indicate almost perfect/significant agreements, and in describing- task as %76 which illustrates a substantial agreement.

5. Discussion

To answer the proposed research questions, it is important to consider the following points: to address the first and second research questions, ANOVA analyses revealed that there were not any significant differences among the analytic and holistic ratings of the tasks. In essence, the type of tasks did not have any effects on the analytic and holistic ratings of students' speaking scores. The inter-rater reliability estimates also showed an almost perfect agreement between the holistic/analytic ratings of two sets of raters for each of aforementioned task. To answer the third research question, the Pearson Product Moment Correlations were conducted. Such analyses indicated a significant correlation between holistic and analytic ratings of the problem-solving, picture-describing and story-telling tasks. However, there was no significant correlation between holistic and analytic ratings as far as explaining task was concerned.

According to Reinders (2009), the role of tasks in supporting second language learning and teaching has been developed recently in many studies (Branden, 2006; Branden, Gorp and Verhelst, 2009; Ellis, 2003; Willis and Wills, 2007). Though there is no study that dealt with the effects of task types and rating methods on learners' speaking scores in an Iranian EFL context. The main object of the present study was to empirically investigate the effects of "task types" including explaining, problem-solving, story-telling and picture-describing tasks and "rating methods"; "holistic vs. analytic ones" on the performance of Iranian EFL learners. Based on the research purpose, participants' performance was analyzed holistically and analytically in terms of pronunciation, vocabulary, cohesion, organization, and grammar. This study found that task types did not have any significant effects on learners' speaking scores and rating methods did not have any effects on students' speaking ratings in problem-solving, explaining, and picture-describing tasks either but it had some effects on students' speaking ratings in the story-telling task.

The result of examining the first research question indicated that the type of tasks did not have any significant effects on the analytic and holistic ratings of students' speaking ability. Such finding is in line with the results of earlier speaking performance assessment, indicating that there was no significant difference in the subjects' holistic rating scores for the three task types, including answering-questions, picture-description, and presentation (McCutchen, 1986; Teng, 2007). Nevertheless, such finding is not in line with the study of Knoch, Macqueen and O'Hagan (2014) that examined to prove whether there is any difference between the discourse made in answer to the independent and integrated writing tasks. The study resulted that the discourse made by the learners' changes significantly on

the most variables under study. In addition, Lim and Lee (2015) showed that the decision-making task group had a tendency towards producing more frequent, but shorter utterances than the opinion exchange task group. Moreover, the results confirmed the effects of both modality and task types on conversational modifications. The participants implemented more meaning negotiation mechanisms in the F2F method and on convergent tasks.

There have been some studies concerning the effects of task types on performance. The findings are inconclusive however. There are some contradictory results in this area though. For example, Blake (2000) confirmed that jigsaw tasks led to better performance, whereas Smith (2003) claimed that decision-making tasks drew out better performance comparing to jigsaw tasks. Lee (2008) found that learners used a higher amount of self-repair in the divergent tasks comparing to that of the convergent tasks. Similarly, Jackson (2011) proved that learners performed better in the divergent tasks. In contrary to Pica, Kanagy, and Falodun (1993) hypothesis, the divergent tasks in these studies induced more satisfactory results. Therefore, the effects of task types need to be investigated more carefully to tackle the inconclusive results. Contradictory findings suggest that further research should be conducted to shed light on the precise relationship between task types and learners' performance.

The investigation of the second research question showed that there was no significant difference between holistic and analytic ratings of the problem-solving, explaining, and picture-describing tasks. However, a significant difference was found between holistic and analytic ratings of the story-telling task. The holistic rating scores were higher than those of analytic ones. These findings are to some extent in line with the findings of Huang (2007) that investigated the effect of task types on the EFL learners' performance on speaking test in Taiwanese context. He concluded that there was no significant difference in the participants' holistic scores for the three task types including presentation, picture-description and answering-questions. However, on the two analytic measures, i.e., fluency and complexity for task types, there were significant major impacts. However, the findings of the most studies concerning comparison of holistic and analytic ratings are in favor of analytic rating method. For example, Saeidi and Rashvand (2012) investigated the impact of task types and rating methods (holistic vs. analytic) on the writing scores of EFL learners in an Iranian context. The results showed that the scores of learners in holistic rating method were higher than their scores in analytic rating method. They claimed that rating methods (holistic vs. analytic) had the significant impacts on learners' writing scores. Harsch and Martin (2013) suggested an integral approach,

combining holistic scores with analytic ones via descriptor-focused scores. The findings of Hunter, Jones and Randhawa (1996) and Chi (2001) were in favor of analytic scoring as well.

Regarding the findings of the third research question, it could be argued that the higher the holistic scores, the higher the analytic scores would be for the problem-solving, picture-describing, and story-telling tasks. While the reverse was true for the explaining-task. The findings of the present study contribute to several implications for language teachers and learners within the classroom context, as well as curriculum and test designers. In addition, this research demonstrates the appropriateness of this integrated approach, holistic scores with analytic scores, by a successful application in speaking assessment. These findings would pave the way for the test-developers/teachers to adopt the most appropriate rating methods for learners' speaking ability.

6. Conclusions

One of the questions of the present study was concerned with the existence of any effect of task types on learners' speaking scores. To find the answer to this question, two one-way repeated-measures ANOVAs analysis revealed that there were not any significant differences among both the holistic and analytic ratings of aforementioned tasks**.** This finding may be due to the similar characteristics of these four tasks. We can thus conclude that the kind of tasks did not have any effects on both the holistic and analytic ratings of students' speaking ability. The type of tasks is not the important factor in speaking performance and characteristic of these four tasks cannot change learners' success in speaking scores considerably. However, the results of this study cannot be taken as evidence for other kinds of tasks.

In addition, it was also intended to determine whether rating methods have any effects on Iranian EFL learners' speaking scores. The results of four paired sample T-tests revealed that the rating methods did not have any effects on students' speaking scores in problem-solving, explaining, and picture-describing tasks either but it had some effects on students' speaking scores in the story-telling task. The research provides an insight into the decision on the implementation of both the holistic and analytic scoring schemes for assessing learners' speaking performance.

Moreover, the third purpose of this study was to investigate whether there is any correlation between rating methods and task types and their effects on the learners' speaking scores. The

results of correlation coefficients showed that there was a significant correlation between holistic and analytic ratings of the problem-solving, picture-describing, and story- telling tasks. There was no significant correlation between holistic and analytic ratings of the explaining task, however. Hence, it can be confidently stated that the holistic ratings of tasks can determine the analytic ratings of tasks and vice versa. In fact, the holistic ratings can be regarded as one of the important factors in determining the analytic ratings in speaking tests and vice versa. It was observed that the higher the holistic scores are, the higher the analytic scores will be on the speaking test.

Due to the limitations of the current study, some of the other aspects of the issue at hand were not fully covered by the researchers. This study investigated the effects of four task types including explaining, problem-solving, story-telling and picture-describing tasks on learners' speaking scores. It should be suggested that a future research taking the effects of other task types with different characteristics on learners' speaking scores seems necessary.

Moreover, in this study, the researchers limited the activity by including only forty participants. Therefore, a larger sample of students from multiple institutions and/or different levels of English Language Proficiency are therefore recommended for future research. Besides, the findings of this study are restricted to an Iranian context. Hence further research seems necessary to see to what extent the results would be changed if the study is replicated in other contexts. In addition, some other variables that are likely to affect the quality of mediation including gender, experience, background knowledge, topic familiarity, level of difficulty and test wiseness are worth heeding. These are the further variables that should be taken into account by other researchers while conducting/replicating the empirical research as they would change the results significantly.

References

Airasian, P. W., & Russell, M. K. (2001). *Classroom assessment: Concepts and applications*. New York: McGraw- Hill.

Branden, K. (2006). Task-based language education: from theory to practice. *Library of Congress Cataloging-in-Publication Data*, *54*(4), 1027-1050.

Branden, K., Gorp, K., & Verhelst, M. (2009). *Tasks in action: Task-based language education from a classroom-based perspective*. UK: Cambridge Scholars Publishing.

Brindley, G. (1994). Task-centered assessment in language learning programs: the promise and the challenge. In Bird, N., Falvey, P., Tsui, A., Allison, D. and McNeill, A. (eds.), HongKong. *Language and learning*, *32*(2), 73-94.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programs: a review of the issues. *SAGE Journals*, *15*(1), 45-85.

Bygate, M. (1999). Task as context for the framing, reframing and unframing of language. *System, 27*(1), 33-48.

Bygate, M., Skehan, P. & Swain, M. (2001). Effects of task repetition on the structure and control of oral language. *Researching pedagogic tasks: Second language learning, teaching and testing, 27*(11), 23-48.

Candlin, C.N. (1987). Towards task-based language learning. In Candlin, C.N. & Murphy, D.F. (eds.). *Language Learning Tasks, 15*(6), 5-22.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of applied measurement, 2*(4), 379-388.

Colomar, A. M. P. (2014). A classroom-based assessment method to test speaking skills in English for Specific Purposes. *Language Learning in Higher Education, 4*(1), 9-26.

Eckes, T. (2005). Examining Rater Effects in TestDaf Writing and Speaking Performance Assessments: A Many Facet Rasch Analysis, *Language Assessment Quarterly, 2* (3), 197-221.

Ellis, N. C. (2001). Memory for language. In Robinson, P. (ed.), *Cognition and Second Language Instruction.* Cambridge: Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2006). The methodology of task-based teaching. *Asian EFL journal, 8*(3), 19-45.

Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus Analytic Evaluation of EFL Writing: A Case Study. *English Language Teaching, 8*(7), 225.

Han, T. & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish Perspective. *Journal of Language Teaching and Learning, 53*, 112-147.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20* (3), 281- 307.

Huang, H. C. (2007). Lexical context effects on speech perception in Chinese people with autistic traits. *Journal of child language, 27* (1), 3-42.

Hughes, J. (1989). Why functional programming matters. *The computer journal, 32* (2), 98-107.

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, *11*(2), 61.

In'nami, Y. & Koizumi, R. (2015). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33* (3), 341-366.

Jackson, S. (2011). *Social works: Performing art, supporting publics*. Routledge: Routledge Publications.

Kim, Y.J., Payant, C., & Pearson, P. (2015). The Intersection of Task-Based Interaction, Task Complexity, and Working Memory: L2 Question Development through Recasts in a Laboratory Setting. *Studies in Second Language Acquisition, 37* (3), 549-581.

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® writing test. *ETS Research Report Series, 2014* (2), 1-74.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131-166.

Lim, C., & Lee, J.H. (2015).The effect of task modality and type on Korean EFL learners' Interactions. *Journal of Asia TEFL, 12 (2),* 87-13.

Long, M. (2015). Second language acquisition and task-based language teaching. *Applied Linguistics, 37* (3), 438-441.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415-437.

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

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language*, *25,* 431-444.

McGraw, H. (2015). 6 types of assessment of learning. *Teach Thought Staff, 24*, 517.

Park, Y., & Park, G. (2004). A new method for technology evaluation in monetary value: procedure and application. *Technovation*, *24*(5), 387-394.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. *Multilingual Matters*, 9-9.

Prabhu, N. S. (1987). *Second Language Pedagogy*. Oxford: Oxford University Press.

Reinders, H. (2009). Study Skills for Speakers of English as a Second Language. *Genome Research, 18*(3), 469-76.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching.* Cambridge: Cambridge university press.

Saeidi, M., & Rashvand, Sh. (2012). The Impact of Rating Methods and Task Types on EFL Learners' Writing Scores. *Journal of English Studies, 1*, 59-68.

Samuda, V. (2001). Guiding relationships between form and meaning during task performance: The role of the teacher. *Researching pedagogic tasks: Second language learning, teaching and testing, 8* (3), 119-140.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education*. US: Cengage Learning.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*(3), 355-390.

Skehan, P. (1998). Task-based instruction. *Annual review of applied linguistics*, *18*, 268-286.

Smith, j. (2003). Reconsidering Reliability in Classroom Assessment and Grading. *Educational Measurement Issues and Practice*, *22*(4), 26-33.

Swain, M. (1995). *Three functions of output in second language learning. Principle & practice in applied linguistics*, Oxford: Oxford University Press.

Teng, H. C. (2007). A Study of Task Types for L2 Speaking Assessment. Retrieved August 25, 2017 from WWW. eric.ed.gov.

Teng, H. & Huei-Chun, S. (2007). A study of Task Types for L2 Speaking Assessment, *Institute of Education Sciences*, 11.

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

Upshur, J.A. & Turner, C.E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16* (1), 82-111.

Willis, D., and Willis, J. (2007). *Doing Task-based Teaching.* Oxford: Oxford University Press.

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing, 2* (1), 59-92.

Appendix

Task Types

1-Problem-solving

a. Think of a town center where there is too much traffic. Think of three alternative solutions to this problem. List the Advantages and disadvantages of each alternative. Then decide which alternative would be the cheapest, the most innovative, and the most environmentally friendly one.

b. You are on a committee that is in charge of deciding what to do with a small amount of money that has been donated to improve your school. You only have enough money for 5 items. You must therefore reach a consensus how to spend the money.

2- Picture-describing

a. Describe a photo or a picture. Each student is given a photo or a picture and having her describe what it is in the photo or the picture.

b. Describe to one of your close friends what happened during the birthday party she had missed due to a severe headache.

3-Explaining

a. Your patient is nervous about her surgery tomorrow. She doesn't understand what is going to happen. Explain to this patient what is going to happen before surgery (e.g., is anything going to happen to prepare her for surgery? Will she have any tests? What is the doctor going to do? Will she be able to eat and drink normally?

b. Explain to your parents why you were so rude in the party yesterday

4- Story- telling

a. Tell me about a special event in your life.

b. Tell me about a perfect day in your life.

c. Tell me about a birthday you remember.

d. Tell me about a time when you gave someone a surprise.