The Effect of Task Types on Raters Assessing EFL Speaking Performance

Tuba İNAN
Istanbul Medeniyet University
Selma KARABINAR
Marmara University

Abstract

Speaking has been one of the most difficult skills to be assessed because the cognition of the raters is complex and uncontrollable as it takes place in raters' minds. The study investigates what goes on in raters' minds during the assessment of speaking performance and potential effects of different task types on raters. Six non-native instructors of EFL were asked to score performances of six examinees on three different oral tasks. Then, the raters produced retrospective verbal reports on how they rated the examinees' performances. The findings revealed that the raters not only attended to the criteria given in the rating scales but they also brought their own criteria into the assessment process. Those non-rubric criteria for each task type included different aspects of assessment such as rater, examinee or task accomplishment. The raters had some concerns over the picture description task while favoring the extended monologue and question and answer task.

Keywords: Rater cognition, speaking test tasks, retrospective verbal reports

Introduction

With the influence of communicative language testing which arose in 1980s, performance-based assessment has been used prevalently in second language speaking assessment (McNamara, 1996; Fulcher, 2000; Brown, 2004). In performance-based assessment, examinees are expected to perform certain tasks (McNamara, 1996). Those performances are assessed by human raters through a rating scale. This means examinees' performances are not the only variable leading to the score. As being different from traditional assessment, performance-based assessment involves more variables such as the rubrics, raters and task types which in turn affect the validity and reliability of the decisions to be made.

Among the variables involved in the oral assessment process, task types could be considered as the key variable that affects the other three: examinees, rubrics and raters. As regards the examinees, tasks could be considered as the guide that provides the context and general format

of the target oral performance during the assessment process (Brown & Yule, 1983; Louma, 2009). However, inherent features of task types may cause variations in the performance of examinees since the relationship between the examinee and the task, and the way each examinee deals with the task might be different (Bachman, 1990; McNamara, 1996; Skehan, 1998; Robinson, 2001).

Task types are also closely associated with rubrics used in the assessment process. In performance-based assessment, tasks help generate open-ended, direct and authentic responses which cannot be marked by using a simple answer key (McNamara, 1996; Upshur & Turner, 1999; Luoma, 2009). Instead, performances have to be evaluated by the use of rubrics prepared according to the specific characteristics of the target task types (e.g. paired or individual, monologue or interview, description or narration etc.) (Stevens & Levi, 2005; Reddy & Andrade, 2010). In the assessment process, rubrics provide a sampling of behavioral domain, which make the features of the target task explicit (Fulcher, 2003). That is why task-specific rubrics have to be written for individual tasks (Brookhart, 2013).

Another important effect of task type is on raters, the other human party involved in the assessment process. Raters do not simply process what the examinees present to them, but they have to engage in highly complex mental processes such as observing, interpreting, recalling information from memory storage and then organizing, combining, weighing and integrating the information to make inferences about examinees (Myford & Wolfe, 2004; Tavares & Eva, 2013). Among those processes, observing seems to be the easiest since the assessment tasks are expected to establish *the constancy of the elicitation input* so that raters are presented similar performances by the examinees (Brown & Yule, 1983). However, raters have to interpret the rubrics at the same time while having different mental representations about specific tasks, which may distort their understanding and application of the given criteria (Bachman, 1990; Myford, 2012). Raters might even have self-generated criteria for different tasks that are not given in the rubrics, which also jeopardize the reliability of their judgment (Campbell, 1993; Meiron, 1998; Brown, 2000). As a result, integration of information in rubrics with what examinees presented to them throughout their performance on the given task could be challenging for raters. All in all, it could be concluded that task type is an important factor affecting not only the examinees' performances but also the content of the rubrics and the judgment of the raters.

Earlier research on the task effects revealed enlightening but contradictory results. To begin with, Brown, Iwashita and McNamara (2005) studied rater orientations in the assessment of five TOEFL speaking tasks: two independent speaking tasks, two integrated tasks. They collaborated with ten trained English for Academic Purposes (EAP) raters without any criteria or scales because the study aimed to find out unguided orientations of the raters. The results showed that raters showed variability in terms of the criteria aimed at different task types since raters seemed to attend differently to two criteria which are *content* and *linguistic resources* depending on task-type.

Chalhoub-Deville and Wigglesworth (2005) also studied the effects of task types on rater judgment. They conducted their study on native speakers of English acting as raters from different nationalities during rating process of ESL examinees' performances on three tasks of Test of Spoken English (TSE): "give and support an opinion", "picture-based narration" and "presentation". The results revealed a statistical difference among the groups in the scores given to examinees' oral performances for each of the three tasks, which means that task type affected ratings.

Similarly, Ahmadi and Sadeghi (2016) also revealed significant task effects on rater judgment in their study where they explored the effect of three different tasks: monologue, interview with the interviewer, and group oral test. The raters were four MA TEFL students with teaching experience but no training. They rated the performances both holistically and analytically. The results of holistic and analytic scoring showed that raters gave different rating when they used different rubrics. It was also found out that the test format had a significant effect on the examinees' scores.

However, as opposed to aforementioned studies, there were some other studies which found out no task effects on rater judgment. Teng's study (2007), for instance, investigated the effect of task types on rater cognition in the assessment of EFL speaking performance and found no significant difference in the examinees' scores. Two trained raters participated in the study and were asked to assess the examinees' speaking performances based on three task types: answering questions, picture description, and presentation by using a holistic scale.

In a similar attempt, Baturay, Tokmak, Doğusoy and Daloğlu (2011) investigated the effects of three different task types which were all picture-cued: narrative, descriptive and prediction-personal reaction. Four teachers evaluated the examinees' performances through a rating scale. Similar to Teng's (2007) study, the results showed no significant difference between the scores given by the four raters. Moreover, there was no significant difference between the separate categories of the rating scale *(grammar and vocabulary use, pronunciation accuracy, fluency and thematic relevance)*.

To sum up, considering the contradictory findings, it could be expressed that there is not enough evidence to reach conclusions about the effects of the task types on rater cognition. In this respect, the current study aimed to investigate whether the task types might have an effect on rater judgment in the assessment of speaking performance. By doing so, it aimed to contribute to the growing body of research into the rater cognition field, especially in the Turkish context. Also, it was observed that the studies conducted on the task effects on rater judgment in the assessment of oral performance are mostly quantitative in their nature. Thus, by adopting primarily a qualitative approach in order to uncover the raters' thinking processes, it aimed to provide a deeper understanding about how task types might have an effect on the rater judgment.

All in all, the current study set out to investigate the following research questions:

1. How do raters' comments about rubric-related and non-rubric factors change according to 3 different task types (question & answer, picture description and extended monologue) while scoring examinees' speaking performance?
2. How do raters' scores change across three different task types?

Method

The present study set out to investigate the decision-making processes of raters during the assessment of speaking performance on three different speaking tasks. To understand what goes on in raters' minds, multi case research design was employed. In a simulation of real-life oral-exam situations, thick descriptions of six raters' retrospective verbal reports were analyzed (Yin, 2003; Gerring, 2007). The qualitative data gathered from retrospective verbal reports were compared with the raters' scores.

Setting

The study was conducted at the English Preparatory School of a state university in Turkey. At the target university, the departments which use English language as the medium of instruction in the undergraduate courses require their students to take a one-year-intensive English preparatory program offered by the School of Foreign Languages. In an academic year of this preparatory programme, oral skills are assessed through speaking exams. In each speaking exam at B1 level, a different task type is applied: the *question and answer* task is conducted with an interlocutor teacher while the *picture description* task and the *extended monologue* task do not require any interlocutor. Each of these tasks is assessed with an analytic rubric prepared by testing unit of the target university (Appendix A). In the application of the speaking exams, two instructors take part as the raters. The average of these two scores is assigned to the examinees.

Participants

There were two groups of participants in this study: six raters (2 male and 4 female) and six examinees (3 male and 3 female). The raters were all from the same educational (both BA and MA in ELT) and professional backgrounds (with 5-10 years of teaching and testing experience). Performing the three speaking tasks was the only contribution of the examinees. The examinees were selected based on their convenience and willingness; however, their English levels and departments were kept constant. All the examinees were medical students at B1 level.

Data Collection and Analysis

Overall design of the study had three phases. In phase I, demographic information of the examinees were collected and their performances were video recorded. The second and the third phases of the study included only the raters. The second phase involved having demographic information of the raters and consent forms, recording their retrospective verbal reports and the scores that they assigned to the examinees' performances. The raters were asked to watch the examinees' performances based on the tasks in the following order: the *question and answer*, the *picture description* and the *extended monologue*. In order to eliminate the possible sequence effect (Plous, 1993; Attali, 2011), each rater assessed the performance of the examinees in a different order. Besides, when the raters were producing their retrospective verbal reports, the researcher took field notes so as to shape the questions to be asked in the semi-structured interviews. In the last phase, in order to clarify the information provided in the raters' retrospective verbal reports and the researcher's notes, semi-structured interviews with the raters were carried out.

Speaking Tasks

In order to simulate the real-life exam procedure and minimize the material and topic effect, 2 sets of tasks were utilized. In each set, there were 3 task types (question and answer, picture description and extended monologue). The tasks in each set were similar but not the same which were all based on the topics derived from the curriculum. The examinees were assigned to one of the sets randomly and were asked to perform these tasks while their performances were video recorded.

The question and answer task: In the *question and answer task*, the examinees talked to one of the raters who acted as the interlocutor. The interlocutor's only duty was to ask the questions to the examinees without interacting with the examinees unless they requested clarifications.

After a short warm up, the examinees were required to answer 6 questions included in each set and asked by the interlocutor (Appendix B). The examinees were neither allowed to see the questions nor given any preparation time before the exam. The questions were about different topics and were formed to elicit personal answers with some details or examples in at least two or three utterances.

The picture description task: In the *picture description* task, each examinee was given a pair of pictures which were taken in similar contexts. In one of the picture set, there are 2 different types of restaurants (one traditional and one fast food restaurant) to be compared. In the other picture set, there are two different groups of people spending time together in their own ways. The examinees were asked not only to describe both of the pictures in detail but also to compare them. The interlocutor only showed the pictures to the examinees, she did not interact with the examinee. Before the examinees started their performance, they were allowed to have a look

at the pictures for about two minutes so that they could analyze the details that they might use during the description and the comparison process.

The extended monologue task: In the *extended monologue* task, the examinees were expected to express their opinions based on a given topic by producing a two-three-minute speech. For that purpose, each student was asked whether they agreed or disagreed with one of the following statements:

1. Women are less interested in sport than men. Why (not)?
2. Students should be allowed to use their mobiles in class. Why (not)?

After the topics were assigned to students, they were given about 4-5 minutes to get prepared for their performances, and write some notes. During their performances, the students were allowed to use their notes, but they were not interrupted or prompted even if they had difficulty in extending their speech.

Three different analyses were applied to the data obtained in the study: verbal protocol analysis (VPA) followed by frequencies of the comments made by the raters for each tasktype under rubric related and non-rubric factors, content analysis (CA) of semi-structured interviews and researcher's field notes, and calculation of average scores assigned by the raters.

In order to analyze the raters' verbal reports, first audio-recorded data were transcribed and then, following Green's (1998) guidelines, segmented into meaningful units which consist of one or more utterances with a single aspect of the event as the focus. After each segment was assigned with a relevant code, a coding scheme was designed by the researcher and the reliability of the coding was checked by a second coder who was an EFL instructor with a PhD degree in ELT. The second coder was provided with a list of codes that were identified by the researcher in the verbal reports of two raters (Rater 2 and Rater 3). The second coder both segmented and coded both of the verbal reports independently. The inter-coder reliability coefficients were calculated for each verbal report separately by applying the formula provided by Miles and Huberman (1994). Reliability coefficient calculated for the verbal protocol analysis of Rater 2 was 0.91 and of Rater 3 was 0, 93.

Results and Discussion

Changes in Raters' Comments Across Task Types

According to the results of the verbal protocol analysis, the raters attended to not only the criteria given in the rating scales (54,5 %) but they also brought their own criteria into the assessment process (45,5%). The distribution of *rubric-related* and *non-rubric factors* are presented in Table 1.

Table 1

Distribution of Comments about Rubric-Related and Non-Rubric Factors Across Task Types

	Core theme	Task 1 – Question and Answer (*f*)	Task 2- Picture description (*f*)	Task 3- Extended monologue (*f*)	Total *f* (%)
Rubric-related factors	Linguistic factors	142	138	131	411
	Task-specific factors	62	52	43	157
Rubric-related factors Total		**204**	**190**	**174**	**568 (54,5 %)**
Non-rubric factors	Examinee-related	28	12	14	54
	Task-related	42	43	75	160
	Rater-related	70	102	89	261
Non – rubric factors Total		**140**	**157**	**178**	**475 (45,5 %)**
TOTAL		344	347	352	1043

As shown in Table 1, there were two core themes under *rubric-related factors: linguistic factors* and *task-specific factors*. On the other hand, under *non-rubric factors,* there were three core themes which were *examinee-related*, *task-related* and *rater-related*. Those core themes included the codes which were not presented on the rubrics given to raters but still attended by the raters while they assessed examinees' oral performances. In the assessment of *the question and answer task*, *rubric-related factors* were attended the most. This might be because of the fact that examinee performed on *the question and answer task* much longer than they did on the other tasks. Also, contrary to the other two tasks, the raters made more comments on *non – rubric factors* related to *the extended monologue* task (*f*=178). This might imply that the rubric assessing *the extended monologue* task was not as comprehensive as the rubrics assessing the other tasks, or the raters were affected more by *non-rubric factors* for some other reasons while assessing *the extended monologue* task.

Rubric-related Factors

Linguistic factors: *Linguistic factors* contained three sub-themes: *grammar*, *vocabulary* and *pronunciation*. These sub-themes contained several codes such as *grammar mistake*, *vocabulary range* or *intelligibility* which were presented in the analytic rubrics given to raters.

Table 2

Distribution of Rubric-Related Comments Across Task Types

	Core theme	Sub-theme	Task 1 – Question and Answer (*f*)	Task 2- Picture description (*f*)	Task 3- Extended monologue (*f*)	Total *f* (%)
Rubric-Related Factors	**Linguistic Factors**	Grammar	40	51	43	134
		Vocabulary	55	46	40	141
		Pronunciation	47	41	48	136
TOTAL			142	138	131	411

As Table 2 shows, among the three tasks, the *picture description* task contained the most number of comments related to *grammar* (*f*=51) and the least number of comments related to *pronunciation* (*f*=41). This was probably due to the fact that the raters paid more attention to how the language was used rather than how accurately or intelligibly it was pronounced. On the other hand, *vocabulary* was commented on the *question and answer* task more (*f*=55) than the other tasks. This might be probably because of the fact that there were six questions in this task. Each question was about a different topic, which means the examinees were expected to use a wide range of vocabulary items to succeed in the task. In terms of *pronunciation*, there was not a noteworthy difference in any tasks.

The most important task effect on rater judgment in terms of *linguistic factors* was the emergence of *task-specific grammar* expectation. This expectation emerged only in the assessment of the *picture description* task. An example comment made by Rater 6 related to *task-specific grammar* shows that in the assessment of the *picture description* task, some raters expected examinees to use certain grammatical structures such as "present continuous tense" while describing a picture. However, the rubrics did not have any descriptors clearly saying or implying the use of such structures.

Rater 6: "*Now there are some certain things to be considered in evaluating picture description. For example, did they use there is/are or present continuous tense? These were important. The examinee used all of these and she did not make any mistakes in using them, I gave her full points in grammar.*"

Non-rubric Factors:

Examinee-related factors

Examinee-related factors included two sub-themes: *examinees' behavior* and *examinees' language use*. *Examinees' behavior* involved the criteria targeting how examinees behaved during their performances on three task types. This sub-theme emerged because the raters turned out to be affected by the examinees' *self-confidence, calmness, motivation, positive attitude, effort* or *natural speech* during the exam. *Examinees' language use*, on the other hand, involved the criteria in raters' minds related to conversational skills such as how well the examinees *use conversational fillers* and *self-correct*, what *the raters' holistic impression* about the examinees is or what *the examinees' general proficiency level* is and whether the examinees *repeated the mistakes* that they made.

Table 3

Distribution of Examinee-Related Comments Across Task Types

	Core theme	Sub-theme	Task 1 – Question and Answer (*f*)	Task 2- Picture description (*f*)	Task 3- Extended monologue (*f*)	Total *f* (%)
Non-rubric factors	Examinee-related	Examinees' Behavior	16	1	5	22
		Examinees' language use	12	11	9	32
TOTAL			28	12	14	54

As presented in Table 3, e*xaminee-related factors* were referred to more in the *question and answer* task (*f*=28) than in the other two tasks. This is probably the consequence of partial interactive nature of this task since the interlocutor asked the questions and supported the examinees when needed. Although there was mostly one-way interaction in this task, it still involved two people interacting to a certain extent. Therefore, while assessing the *question and answer* task, the raters might have empathized with interlocutor and attended to the *examinee – related* criteria more. The sample comment about e*xaminee-related factors* given below shows how Rater 4 was affected by *self-confidence* of an examinee positively;

Rater 4: "*He was quite self-confident during his performance especially about his confidence in pronouncing the words was what saved him…"*

Similarly, the comment below shows how Rater 3 was negatively affected by hearing the examinee's repetitive mistake in the pronunciation of a certain vocabulary item throughout the exam;

Rater 3: "*… the word 'because' was pronounced incorrectly. It was a repeatedly used word in his performance. Hearing the same mistake again and again was disturbing. I cut 0,5 point*."

In contrast, *the picture description* task involved the least number of comments related to examinees. This might result from the fact that the examinees had to deal with materials while carrying out their oral performance. To elaborate, *the picture description* task did not involve human interaction, which made human-related criteria less important. In contrast, in *the question and answer* task, examinees need to talk to an interlocutor, a real person. This means that human – related criteria, especially examinees' behavior, were more salient to the raters in the assessment of *the question and answer* task.

Rater-related factors

Analysis of raters' verbal reports demonstrated that *rater related factors* involved two sub-themes: *rater reflection* and *rater variation*. The raters not only commented on the examinees' performance but also they made some *reflections on themselves*, on the *rubrics*, and on the *success* or *failure of the examinees*. These reflections did not have any effect on the scores assigned by the raters. On the other hand, the codes emerging under *rater variation* were identified when the raters made inter or intra-examinee comparisons (*the sequence effect*), failed to differentiate the distinct criteria (*the halo effect*), avoided assigning extreme scores (*avoiding extremism*) and employed their own criteria which were not presented in the rubrics (*mental rubric*). Comments related to rater variation affected the decisions about examinees' scores.

Table 4

Distribution of Rater-Related Comments Across Task Types

Core theme	Sub-theme	Task 1 – Question and Answer (*f*)	Task 2- Picture description (*f*)	Task 3- Extended monologue (*f*)	Total *f* (%)
Rater-related	Rater Reflection	33	35	41	109
	Rater Variation	37	67	48	152
TOTAL		70	102	89	261

As shown in Table 4, regarding *rater reflection*, there was not a noteworthy difference in the number of comments made for each task. However, in terms of *rater variation*, the comments in *the picture description* task outnumbered the comments made for other tasks (*f*=67). It had almost twice as much frequency as the *question and answer* task (*f*=37). This is probably due to the fact that the *picture description* task was the second task that the raters evaluated, which increased the number *of intra-examinee sequence effect*. Likewise, the reason of low *rater variation* in the *question and answer* task was probably the lower frequency of the *sequence effect* since it was the first task that was evaluated by the raters, which made only inter-examinee comparisons possible. In other words, while evaluating the first performance, it was impossible to make comparison between the task performances of the same examinee. The second reason for the high frequency of *rater variation* in the *picture description* task was the involvement of *mental rubric* and emergence of the *halo effect* which were the highest in the *picture description* task. Sample quotations regarding the *halo effect* and *mental rubric* are as follows:

Rater 6: "*He described the pictures but did not compare them. He said very few sentences. He even did not use present continuous tense while describing the pictures. I did not hear any of the sentences that I had in my mind.*" (*Mental rubric*) (3 points)

The sample comment made by Rater 6 above shows that the rater had certain expectations related to *the picture description* task. She clearly stated that she wanted the examinee to use certain grammatical structures to describe the pictures. That is why she cut points from the *task-specific* part of the rubric and gave the examinee 3 points. However, there was not such a descriptor in the rubric given to the rater. Thus, it could be said that apart from the criteria given in the rubrics, the rater had her own criteria that she applied in her scoring.

Rater 4: "*She only described what they were doing in the pictures. She did not compare. Because her task achievement was not good, her vocabulary score automatically decreased to 4 points.*"(The halo effect)

The sample comment of rater 4 is an example of *the halo effect* that the raters faced in their scoring. While assessing *the picture description* task, Rater 4 stated that the examinee failed in the *task achievement* criteria, but instead of cutting points only on that part, she lowered her

score on the *vocabulary* criterion, too. This means that Rater 4 had failed in attending to each criterion individually, which resulted in *the halo effect*.

Task Related Factors Appear Both in Rubric-Related and Non-Rubric Criteria

Although analytic rubrics had included a *task specific* part related to the three task types, the raters still needed to mention more *task-related criteria* that were not available in the rubrics. As a result, *task-related factors* include comments about both the criteria appear in the *task specific part of the rubric* and the criteria about how examinees were expected to accomplish the task (*task completion)* that were not directly or clearly stated in the rubrics, but the raters had in their minds.

Table 5

Distribution of Comments about Task Related Factors

	Core theme	Sub-theme	Task 1 – Question and Answer (*f*)	Task 2- Picture description (*f*)	Task 3- Extended monologue (*f*)	Total *f* (%)
Task – Related Factors	Rubric related	Task specific Rubric	62	52	43	157
	Non Rubric	Task completion	42	43	75	160
TOTAL			104	95	118	317

As presented in Table 5, the findings revealed that the number of comments about *task – related factors* appear under *non-rubric criteria* (*f*=160) slightly exceeded the number of *rubric-related* comments (*f*=157).The interpretation of it could be that for the raters, the criteria presented in the rubrics were not comprehensive enough for each of the tasks; especially for the *extended monologue task* (*f*=75) because the raters needed to attend to more criteria than they were supplied with the rubric.

After the analysis of task related comments about the criteria both in the rubrics and in raters' minds, it was observed that different task types not only share some common characteristics, but also they have distinct features. In the assessment of all the three tasks, raters paid attention to common characteristics which were already included in the rubrics such as whether the examinees *hesitated, paused, needed support or prompting* or *repeated* what they said during their performances. Raters paid attention to some more common characteristics which were not included in the rubrics but the raters thought as important such as the *amount of data* the examinees supplied, whether the examinees *presented relevant ideas, conveyed their message*, or how *fluent* the examinees' speech was. Also, all the raters *reflected on the examinees' performance in general*.

On the other hand, it was found out that while assessing *the extended monologue* task, the raters attended to certain distinct criteria. Among those, *rubric-related* criteria specific

to the *extended monologue* task were whether the examinees used *cohesive devices*, how *comprehensible* their speech was and how well they *organized their thoughts*; *non-rubric* criteria were how much they talked *off topic*, whether there was *content repetition*, whether they *read from the notes* and whether they had problems with the *time limit* by exceeding or speaking less than the given time period.

Changes in the Raters' Scores Across Task Types

In order to answer the second research question which investigated the changes in the raters' scores while assessing the three task types, mean scores assigned by each rater and the raters' semi-structured interviews were used.

Table 6

Average Scores Assigned by the Raters to each Task Type

	Question and answer	Picture description	Extended monologue
Rater 1	18,3	14,6	15,2
Rater 2	16,4	17,6	15,1
Rater 3	17,5	17	17
Rater 4	18,5	17,3	16,9
Rater 5	15,8	15,7	15,8
Rater 6	15,3	13,3	13,8
General average	17	15,9	15,6

*The scores were given out of 20 points.

Table 6 presents the mean scores assigned to each task type by the raters. Despite the inter-rater and intra-rater fluctuations, it could be clearly seen that the highest individual rater's mean scores (rater 4=18,5 and rater 1=18,3) and the highest mean score of all raters (17) belonged to *question and answer* task. This finding could stem from several reasons. First of all, in their semi-structured interviews, two of the raters expressed that the *question and answer* task was the best for assessing speaking performance. This might have affected their scores in a positive way and they might have had a tendency to award higher scores.

In the sample quotation obtained from his semi-structured interview, Rater 2 expressed why the *question and answer* task was the best task to assess speaking performance:

Rater 2: *"I think question and answer is better because it involves communication. If the questions are chosen appropriately, it is the best task to assess oral performance."*

Another reason for the higher scores assigned to the *question and answer* task might be that the examinees' performances on the *question and answer* task were far longer than the other tasks, which might have supplied more data to the raters in favor of the examinees. The length of the examinees' performances might have given the impression that the examinees were better on that task.

The results also showed that the average scores assigned to the *picture description* task and the *extended monologue* task were close. This might be interpreted in different ways. To begin with, examinees might not have performed on both of these tasks as well as they did on the *question and answer* task. This means that these tasks somehow affected the examinees' performance or the raters' scoring negatively. Secondly, although raters expressed their concerns over the *picture description* task in their semi-structured interviews and their retrospective verbal reports, their prejudice was not reflected in their scores. For example, Rater 4 was one of the four raters who favored the *extended monologue* as the best task to assess oral performance while she criticized and expressed her concern over the *picture description* task. However, despite her concerns over *picture description* task, she had the second highest average score (17,3) assigned to the *picture description* task.

The comment taken from semi-structured interview with Rater 4 can exemplify her point over these two tasks:

Rater 4: "*I think the best one is extended monologue. We can see what the examinees have in their minds. There is a lot to talk about in the questions. They have the opportunity to extend their answers. Picture description limits the examinees even more. They cannot extend their answers. We cannot see their real potential.*"

Conclusion and Implications

The study revealed that the raters attended to not only *rubric-related* but also *non-rubric factors* while assessing all the three tasks. In other words, in addition to the specified criteria, the raters considered aspects of performance that were not explicitly included in the rubrics. It can be concluded that raters tend to bring their own criteria into assessment process in all three task types used, which is quite similar to what was referred to as self-generated criteria by Meiron (1998), and some other study findings in the related literature such as Brown (2000) and Brown et al. (2005).

First of all, by combining two research questions, conclusions can be made about each target task type, namely the *question and answer*, the *picture description* and the *extended monologue*. To begin with, it was found out that the raters preferred using *question and answer* task for assessing speaking skill because it had an interactive nature and somehow enabled the examinees to talk longer. Perhaps, as a result of it, the raters assigned highest scores to performances of this task while paying more attention to *examinee – related* criteria such as *self confidence* and *mistake repetition*.

Regarding the *picture description* task, there were several conclusions. Firstly, the raters criticized the task as having a non-interactive nature and they did not comment on the *examinee – related* factors much. The raters were commonly faced with the *halo effect* and were affected by their own *mental rubric*. However, despite the differences among the raters' expectations on

the *picture description* task and their concerns over it, the raters' scores were not much lower or higher than the scores assigned to extended monologue task. Finally, some raters expected the examinees to use certain *task specific grammatical structures* in the performance of the *picture description* task whereas some of the raters did not consider the use of *task-specific grammar* as a must. Since not all the raters have the same expectations, rubric descriptors should be paid more attention and should include more details regarding the *picture description* task.

As for the findings concerning the *extended monologue* task, it was found out that comments about *non-rubric factors* were the highest in number. Those non rubric criteria were all about *task accomplishment* since perhaps this task has distinct characteristics (such as taking notes and reading from those notes) and they were not clearly included in the rubric. This can lead to a conclusion that *task-specific criteria* given in the rubric does not fulfill raters' expectations; therefore, rubrics for that task needs special care. Surprisingly, although the *extended monologue* task was not interactive, it was not criticized as much as the *picture description* task.

In terms of *rubric-related* assessment criteria in general, it could be expressed that there was an agreement among raters, since there were not noteworthy differences in their understanding of *task-specific* features stated in rubrics. Grounded on the findings, it could be concluded that except for some differences in their interpretation of certain criteria related to what is expected from *picture description* task (mentioned above), the raters were more or less in the same path in the definition of the speaking skills since they attended to similar criteria while assessing all three tasks. This finding could be considered in parallel with the study conducted by Chalhoub-Deville and Wigglesworth (2005) who concluded that native speaker raters from different nationalities agreed on the definition of the speaking construct. This finding is promising in terms of construct validity of the speaking skill. Based on the assumption that what raters commented in their verbal reports is actually the representation of the speaking construct in their minds (Davies et al., 1999), different speaking tasks do not make significant changes in their definition of speaking construct. To sum up, within the limitations of this small scale study, it can be concluded that while each task type included in this research may have some certain characteristics that are specific to only that task itself, there are still common characteristics that are shared by all three task types.

Regarding the studies in the related literature, there were contradictory results among the ones conducted statistical analysis to investigate rater variation in scores depending on different task types. Teng (2007) and Baturay et al. (2011), for example, found no differences among the raters whereas Ahmadi and Sadeghi's (2016) study revealed significant differences. Chalhoub-Deville and Wigglesworth's study (2005), on the other hand, found a difference in rater judgment concerning different tasks, but the difference was not big enough. As for the current study, according to the mean scores, the raters turned out to have assigned the highest scores to the *question and answer* task. However, it is not possible to draw a conclusion related to the task effects on the scores since no statistical analyses were conducted.

In the light of these conclusions, the current study provides useful implications for test developers, teachers and students. First of all, the study revealed that different test tasks did not create huge differences in raters' minds, but still, raters had concerns over certain tasks. In this respect, test developers should pay more attention to the limitations of the task types. Taken this into consideration, it could be suggested that a variety of task types should be employed for assessing oral performance. Another implication that could also be useful for test developers is that a separate rubric for each oral performance task should be prepared in detail considering the different aspects and requirements of the target task.

Also, there was an important implication for the teachers acting as raters. Differences in raters' scores regarding the same task or across different task types bring about the rater reliability issue. This implies a need for rater training concerning different task types since different task types may create different expectations in raters' minds. By training the raters, they could be familiarized with task requirements and their relation to rubric criteria that they are supposed to use while evaluating speaking performance.

The final implication could be for the students. Since task familiarity is important for the examinees, the students should be given more opportunity to practice speaking tasks in class so that they can perform to their highest potential in the exams. Also, task specifications and the rubrics should be shared with the students before the exams since they need to know the exam requirements.

Regarding suggestions for future research, first of all, potential effects of task types on the examinees' performances can be studied by collecting data from students as some studies in the literature revealed that students performed differently on different task types. Also, studies can be conducted on the inherent task characteristics such as the effect of topic, planning time, task difficulty and task complexity. Finally, replications of this study can be conducted with the involvement of the raters who are native speakers of English, or with the involvement of different task types.

References

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. Language Assessment Quarterly, 13(4), 341-358.

Attali, Y. (2011). Sequential effects in essay ratings. *Educational and Psychological Measurement*, *71*(1), 68-79.

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

Baturay, M. H., Tokmak, H. S., Dogusoy, B., &Daloglu, A. (2011). The impact of task type on oral performance of English language preparatory school students. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education)*, *41*(1), 60-69.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Virginia: ASCD.

Brown, A. (2000). An investigation of the rating process in the IELTS Speaking Module. In R. Tulloh (Ed.), Research reports (1999, Vol. 3, pp. 49-85). Canberra, Australia: IELTS Australia.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test taker performance on English for academic purposes speaking tasks. *ETS Research Report Series*, *2005*(1), i-157.

Brown, G. & Yule, G. (1983). *Teaching the spoken language: An approach based on the analysis of conversational English*. Cambridge: CUP.

Brown, H. D. (2004). *Language assessment, principles and classroom practices*. New York: Pearson/Longman.

Campbell, E. H. (1993). *Fifteen Raters Rating: An Analysis of Selected Conversation during Placement Rating Session*. Speech presented at The Annual Meeting of the Conference on College Composition and Communication, San Diago, CA. 25

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383-391.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing.* Cambridge, UK: Cambridge University Press.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, *28*(4), 483-497. Fulcher, G. (2003). *Testing second language speaking.* Harlow: Pearson Education.

Gerring, J. (2007). *Case study research: Principles and practices.* New York: Cambridge University Press.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press. Luoma, S. (2009). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Meiron, B. E. (1998). Rating oral proficiency tests: A triangulated study of rater thought processes. Unpublished master's thesis, University of California at Los Angeles.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd Ed.). Sage Publications.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice*, *31*(3), 48-49.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Plous, S. (1993). *The psychology of judgment and decision making*. New York, NY, England: Mcgraw-Hill Book Company.

Reddy, Y. M.,& Andrade, H. (2010). A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, *35*(4), 435-448.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Stevens, D. D. & Levi, A. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Stylus Publishing, LLC.

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.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater based assessments. *Advances in Health Sciences Education*, *18*(2), 291-303.

Teng, H. C. (2007, April). *A study of task type for L2 speaking assessment*. Paper presented at the Annual Meeting of the International Society for Language Studies (ISLS), Honolulu, HI.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd Ed.). Thousand Oaks, CA: Sage Publication Inc.

Appendix A – The Rubric

Linguistic part of the rubric (Same for all tasks – when answering questions about self and talking about everyday situations// when describing and comparing pictures// when talking about given topics)				Task-specific part of the rubric (Use related column for each task)		
	Grammar	Vocabulary	Pronunciation	Question and Answer: Comprehension & Ability to respond	Picture Description: Picture description & comparison	Extended Monologue: Comprehensibility & Discourse Management
5	Shows a good degree of control of level appropriate grammatical forms.	Uses quite a lot of level-appropriate vocabulary	Is mostly intelligible, and has some control of phonological features.	Understands the questions accurately and responds appropriately with almost no difficulty. There are a few noticeable pauses, and hesitation. Requires no support such as repetition or paraphrasing while answering.	Both describes and compares pictures with no prompting and support. There are a few false starts and reformulations.	Produces extended stretches of language with little hesitation and pauses. Clear organization of ideas. Uses a range of cohesive devices.
4	Shares the features of both bands 3 and 5					
3	Shows sufficient control of level appropriate grammatical forms.	Uses a range of appropriate vocabulary despite limited control.	Is mostly intelligible, despite limited control of phonological features.	Understands the questions accurately and responds appropriately with a little difficulty. There are some pauses, and a little hesitation. Requires some support such as repetition or paraphrasing while answering.	Both describes and compares pictures with a little prompting and support. There are some false starts and reformulations.	Produces extended stretches of language with some hesitation and pauses. Some repetition. Uses a few cohesive devices.
2	Shares the features of both bands 1and 3					
1	Shows only limited control of a few grammatical forms.	Uses isolated words and phrases.	Has very limited control of phonological features and is often unintelligible.	Has considerable difficulty understanding and responding. Full of pauses and hesitation. Requires additional support such as repetition or paraphrasing while answering.	Has considerable difficulty in both describing and comparing pictures and requires additional support and prompting. Full of false starts and reformulations.	Produces short phrases with frequent hesitation and pauses. Frequent repetition. Uses basic cohesive devices.
0	Performance below 1					

Rater's name	Grammar (5)	Vocabulary (5)	Pronunciation (5)	Task-specific part (5)	Total (20)
Student's Name					

Appendix B – Sample Questions Asked in the Question and Answer Task

- Warm up:
- How are you?
- What's your name?
- What is your department?
- Where do you live/come from?

Interview questions:

1. Is there anything you'd love to buy but you can't afford at the moment? Why do you want to buy that?

2. What's the most difficult thing about studying English? Why?

3 Is there a sport that you really like watching or doing? What is it? Why?

4. Have you ever been disappointed by a birthday present? Why?

5. Where do you live? What do you like about the area where you live? What don't you like?

6. If you had only 24 hours (one day) to live, what would you do?