Journal of
Language and Translation
Volume 9, Number 3, Autumn 2019, (pp.85-99)

Students' Oral Assessment Considering Various Task Dimensions and Difficulty Factors

Houman Bijani[*1], **Adnan Satariyan**[2]
[1] Assistant Professor in Applied Linguistics, Department of English Language Teaching, Zanjan Branch, Islamic Azad University, Zanjan, Iran
[2]TESOL Researcher, College of Arts, Law and Education, University of Tasmania, Australia

Received: 22 May, 2019 Accepted: 10 October, 2019

Abstract

This study investigated students' oral performance ability accounting for various oral analytical factors including fluency, lexical and structural complexity and accuracy with each subcategory. Accordingly, 20 raters scored the oral performances produced by 200 students and a quantitative design using a MANOVA test was used to investigate students' score differences of various levels of language proficiency groups with respect to their oral scores in each analytical factor. The findings showed that students, in each level of language proficiency, were different from each other regarding various measures of fluency, lexical complexity, structural complexity and accuracy when performing the five oral tasks. Besides, the findings showed that language planning, perspective and immediacy were the determining dimensions in oral task difficulty. The findings demonstrated the usefulness of analytical approaches to rater training programs in detecting rater effects and demonstrating the consistency and variability in rater behavior. The analysis confirmed that the nature of second language oral construct is not constant, thus different results are achieved using different oral task dimensions. Consequently, the outcomes have constructive implications in the use of feedback as a reliable indicator of task difficulty and specifically as a basis for test design and validation.

Keywords: Accuracy; Complexity; Fluency; Oral performance assessment; Oral tasks

INTRODUCTION

Task, according to O'Sullivan (2002), is defined as "bounded classroom activities in which learners use language communicatively to achieve an outcome with the overall purpose of learning language" (p. 278). The current popularity of performance assessment has led to a growing interest in tasks as a tool for assessing learner ability. Task-based assessment engages students in the performance of tasks which stimulates the kind of language found in the real world situation with the purpose of eliciting authentic language sample from the students. One issue regarding variation in students' performances attributes to task characteristics. This variation results in different scores under various conditions; thus, making it a feature of interest for further investigation. May (2009) argues that tasks designed for oral assessment should be magnificently short, well-balanced, and provide opportunities for each

[*]Corresponding Author's Email:
houman.bijani@gmail.com

student when more than one student is participating. When discussing the scoring tasks, Huei-Chun (2007) suggests that the task developer should consider the complexity and length of any texts which are to be used, the difficulty of the vocabulary needed to complete the task, the expected speed of speech, the number of speakers, the explicitness of information, discourse structure, and the amount of non-linguistic support available.

The appearance of IRT has made it possible to investigate task difficulty as in isolation from rater severity (Wolfe, 2004). This is based on the assumption that the scores awarded to an individual on a speaking task are influenced by his/her speaking proficiency, difficulty of the task and the severity of the rater(s). In fact, very little is known about task difficulty or the difficulty of various tasks as they are compared with one another. Consequently, one of the most important challenges, which influences task characteristics, is how to determine task difficulty. This can help us in the appropriate use of task ranges which will clarify the way levels of performance are described. Some scholars (e.g., Skehan, 1998) have identified a number of factors that affect task choice. As an example, Skehan has identified a number of factors influencing task difficulty:

1. *Familiar information:* The more familiar information the task contains, the more fluent the students' performances will be.
2. *Structured tasks:* whenever the task has clear and sequenced structure, it will result in greater fluency and accuracy.
3. *Complex and numerous operations:* the more operations and transformations a task is based on, the more difficult and complex it will be. This will result in lower fluency and accuracy.

Skehan further claimed that by the manipulation of these factors, task performance will vary resulting in variation of task quality.

Factors affecting learners' second language task production

There are several factors such as anxiety of the second language (L2) learners, planning time, familiarity with the topic, genre of the tasks, learner's proficiency level, task type, task structure, task condition, and the level of cognitive complexity of the tasks which influence the performance of second language learners; for example, their speed of production and complexity of their utterances (Ling, Mollaun, & Xi, 2014). As Trace, Janssen and Meier (2017) claim, the L2 learner`s performance differs from task to task. So, L2 learner`s production will be different when they perform different task types, and consequently these different types of tasks will result in variation, called "task-induced variation". May (2009) agrees with this variation and asserts that in performing different tasks, learner`s production of some grammatical, morphological and phonological forms will vary in a particular manner. Skehan and Foster (1999) investigated the role of task type in foreign language oral production in terms of accuracy, fluency, and complexity. Two types of tasks (instruction task and an argumentative task) used in the study. Participants in the instruction-task group performed significantly better than those in argumentative-task group in terms of accuracy, fluency, and complexity. The argumentative speeches were produced with more complex language than the instruction ones. Fluency was higher in instruction speeches. In terms of accuracy, instruction-task group performed better than those in argumentative-task group, but argumentative speeches were more accurate than instruction speeches.

The effect of task type on oral production

Oral assessment is often carried out by considering students' ability to produce words and phrases by evaluating their ability in doing a variety of tasks such as asking and answering questions about themselves, doing role-plays, making up minidialogues, defining or talking about some pictures or talking about given theme. As

Robinson (2001) stated, features of second language oral output such as accuracy, fluency and complexity differ by task type. When second language learners speak, the speed of their production, the complexity of their utterances, and the accuracy of their speech is influenced by a number of factors, including their anxiety, their proficiency, or the degree of task complexity, and the pressure of time. These three aspects, complexity, accuracy, and fluency of learners' performance are considered as learners' language ability determining factors (Robinson, 2001). Studies incorporating task has been primarily concerned with analyzing the impact of task design on the accuracy, fluency and complexity of language in oral production.

In'nami and Koizumi (2016) suggested field testing of tasks along with the use of questionnaires to elicit students' and raters' perceptions to determine good and poor tasks. In an attempt, they scaled a number of oral speaking tasks, used in the ACTFL oral assessment, based on their functions. Then, by the use of a Rasch partial credit model, they assessed the difficulty of a number of tasks. Although they found a reasonable correlation between the suggested difficulty level and the assessment of difficulty by raters, this is far from testing tasks on students and assessing task difficulty from scores.

On behalf of language testing and assessment, Elder, Iwashita, and McNamara (2002) claim that the more difficult and complex a task is, the more difficult it will be. In an attempt, they aimed to modify Skehan (1998) model of task difficulty factors through investigating the following criteria: 1. *Perspective:* To tell a story from one's own perspective or from a third person's. 2. *Immediacy:* To tell a story with and without pictures. 3. *Adequacy:* To tell a story from a complete set of pictures, and with two or four missing ones. 4. *Planning time:* To do an oral task with 2-3 minutes planning time, and without it. Skehan and Foster (1999) have suggested that more complex tasks direct students' attention to context and divert it away from form, whereas simple tasks produce more fluent and accurate speech which is on the opposite side of complex tasks which lead to the production off more complex speech. They further investigated the effect of planning time and three various tasks (personal information exchange, narrative, and decision making) on the variables of fluency, complexity and accuracy. The results showed that planning time had more influence on narrative and decision making tasks than on personal information exchange tasks.

In a similar study, Nakatsuhara (2011) investigated the effect of planning time on students' accuracy, fluency, and complexity measures. The study revealed that those students who had planning time had a better performance with respect to complexity (number of subordinations), fluency (number of self-repairs), and accuracy (lack of grammatical mistakes). Ahmadian and Tavakoli (2011) investigated the effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in the oral production of EFL learners in the context of Iran. The results obtained from one-way ANOVAs revealed that the opportunity to engage simultaneously in careful online planning and task repetition enhances accuracy, complexity, and fluency significantly. In another study Kuiken and Vedder (2014) aimed at investigating the impact of planning conditions on the oral performance of the EFL learners while performing structured vs. unstructured tasks. Results demonstrated that planning time served no impact with regard to the accuracy and fluency of the learners' performances, but resulted in more complex performances when participants conducted the unstructured task. In the meantime, task structure did not affect the accuracy and complexity whilst promoting the fluency under the planned condition. Davis (2016) discussed the fact that 1 min of pre-task planning should be considered as an alternative to extend the face validity of the test. Moreover, although pre-task instructions displayed some role for diverting attention to form, planning did not serve any impact. Leaper and Riazi (2014) extended Robinson's research by crossing a complexity variable (Here-and-Now) with a condition variable (open vs. closed). They hypothesized that the Here-and-Now/There-and-Then narrative would be more complex than the other versions of the task. The results showed that learners who performed the most complex versions of the task were significantly less fluent, with no such large differences regarding either structural or lexical complexity, and with significant improvements with regard to error-free units but not target-like use of articles.

In speaking, rater training is used to modify raters' expectations of tasks and students' characteristics (Khabbazbashi, 2017), and to clarify various elements of the rating scale in order to reduce levels of rater variability (Khabbazbashi, 2017). Training is used to reduce extreme differences by minimizing random errors between raters in terms of severity and to increase the self-consistency of individual raters by reducing random errors (Davis, 2016). Closely related to training are the concepts of rater experience or rater expertise. Because scoring second language oral proficiency is done by raters, they are an essential part of proficiency assessment. Therefore, not only does rating reflect students' oral ability, but also raters' assessment schemes (Attali, 2016). A variety of researches on experienced and inexperienced raters' performances have indicated higher inter-rater consistency following training (Attali, 2016; Bijani, 2010). Commonly, in all the studies, extremely severe or lenient inexperienced raters have benefited from the training program thus have modified their rating behavior making it like the other raters'. In a study by Bijani (2010) on the effect of rater training on rater consistency scoring students' written language proficiency, the consistency of inexperienced raters improved much more after training compared to experienced raters.

However, the relative contribution of tasks factors to the success of any given task for the purposes for which it was designed is mostly unknown. Although it is frequently claimed that lack of specialist knowledge about the task topic makes the task difficult for students, there is little evidence in this case. Still the relative measures between test students' general language proficiency and their oral performance ability is not definite. On the other hand, almost all studies so far have investigated tasks separately; therefore, any possible relationship among them has remained unexplored. On the other hand, the notion of task difficulty and its relationship to underlying subcategory measures of fluency, accuracy and complexity on various task dimensions has not been addressed comprehensively and there is little evidence suggesting which tasks are more suitable for students' of particular ability levels. Consequently, the study investigated measures of oral task difficulty in relation to raters' severity, bias and consistency. It was also identified which underlying factors including accuracy, fluency and complexity with the relationship among which would account for the rating of each task Thus, the following research questions could be formed:

RQ: Do students with various levels of language proficiency differ in speaking ability and speech production factors?

METHODS

Participants

A stratified random sample of 200 adult Iranian students of English as a Foreign Language (EFL), including 100 males and 100 females, ranging in age from 17 to 44 participated as test takers. The students were selected from intermediate, high-intermediate, and advanced levels studying at the Iran Language Institute (ILI). In other words, they were selected in a way that they represented three levels of language proficiency based on their class level placements and teachers' reports of their learning history; thus, their speaking ability levels were controlled while other student characteristics such as gender, age, native language, educational background and the number of years of probable residence in English speaking countries were not. The reason for choosing intermediate to advanced learners of English was that these students had already acquired the adequate knowledge regarding the required elements and standards of oral production. Among the many characteristics, the students' speaking ability was what the test intended to measure, and this was what the raters were supposed to focus on while scoring. The students were randomly selected to take a sample TOEFL (iBT) test including listening, structure, and reading comprehension to make sure that they were not at the same level of language proficiency and that there was a significant difference among the three groups.

A convenience sample of 20 Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 participated as raters. These raters were undergraduate and graduate in English language related fields of study, teaching in different universities and language institutes. The raters participating in this study were naturally all proficient but with a variety of levels of expertise; that is, the raters were different in terms of level of teaching, ranging from basic to advanced. It should also be stated that all the raters had high levels of English language proficiency although none was a native speaker of English language. In order to fulfill the requirements of this study, the raters had to be classified into two groups of experienced and inexperienced raters to investigate the similarities and differences among them and the likelihood advantages of one group over the other one; therefore, a background questionnaire, adapted from McNamara and Lumley (1997), eliciting the following information including (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training* and (5) *relevant courses passed* was given to the raters. Thus, raters were divided into two levels of expertise on the basis of their experiences outlined below.

A. Raters who had no or less than two years of experience in rating and receiving rater training, and had no or less than 5 years of experience in teaching and passed less than the 4 core courses related to ELT major.

B. Experienced raters who had over two years of experience in rating and receiving rater training, and over 5 years of experience in teaching and passed all the four core courses plus at least 2 selective courses related to ELT major.

Materials

The elicitation of students' oral proficiency was done through the use of five different tasks including description, narration, summarizing, role-play and exposition tasks. Task 1 (*Description Task*) is an independent-skill task which reflects students' personal experience or background knowledge to respond in a way that no input is provided for it. On the other hand, tasks 3 (*Summarizing Task*) and 4 (*Role-play Task*) reflect students' use of their listening skills to respond orally. For tasks 2 (*Narration Task*) and 5 (*Exposition Task*), the students are required to respond to pictorial prompts including sequences of pictures, graphs and tables. The tasks were implemented via two methods of task delivery: (1) direct and (2) semi-direct. The direct test was designed for use in an individual face-to-face method (i.e., speaking to an interlocutor _ here a rater), whereas the semi-direct test was designed for use in a language laboratory setting.

For the purpose of comparability, both formats of the test consist of one-way exchanges (monologic) in which the student is required to communicate information in response to prompts from the interviewer/rater. However, on the direct version of the test, the role play allows for a more authentic information gap activity in which meaning is negotiated between a student and an interviewer (dialogic). The tasks were also classified as either planned (allowing preparation time) or unplanned (designed to elicit spontaneous language). In the third place, they were distinguished as either open (allowing a range of possible solutions) or closed (allowing a restricted set of possible responses). In the fourth place, the tasks were also classified as being convergent (involving problem-solving in which the aim is to arrive at a particular goal) and those which are divergent (without specific goals, involving decision making, opinion and agreement). In this study, the only two-way task, role-play, is regarded to be convergent. In another classification, tasks were classified regarding perspective dimension. This was to ask the students to do the tasks from their own (first person perspective) or another person's point of view (third person perspective). Finally, tasks were classified regarding their immediacy dimension. This was to ask the students to speak using Here-and-now and There-and-then language structures. In this respect, the task types used in this study could be classified into two categories regarding their difficulty level based on the given factors above (Robinson, 2001). The following table (Table 1) gives the classification of tasks and their predicted difficulty on behalf of the given factors with respect to their difficulty levels.

Table 1.
Table of Predicted Task Difficulty Classification

Dimension	Difficult (predicted)	Easy (predicted)
Openness	Close (limited response)	Open (free response)
Information exchange direction	Dialogic	Monologic
Language convergence / divergence	Convergent	Divergent
Language planning	Without planning time	With planning time
Perspective	3rd person point of view	1st person point of view
Immediacy	There-and-then	Here-and-now

For both versions of the test, each student's task performance was assessed using the ETS (2001) analytic rating scale. In ETS (2001) scoring rubric, individual tasks are assessed using appropriate criteria including *fluency*, *grammar*, *vocabulary*, *intelligibility*, *cohesion* and *comprehension*.

A questionnaire was used to elicit the students' feedback on both versions of the speaking tests through focusing on their attitudes and feelings, effectiveness and evaluation, clarity and further development of the speaking assessment quality. The questionnaire had originally been developed by Luoma (2004) consisting of five items; however, in order to make it more suitable for this study, it was modified thus the new version consisted of 22 and 18 items, for the direct and semi-direct test versions respectively, on a Likert scale to ascertain whether students' reactions differed significantly according to their characteristics.

Procedure

Having made sure that the three groups of students were at different levels of language proficiency and identified the raters' background information and their level of expertise and classified them as inexperienced raters and experienced ones, the speaking test started with a practice question so that the students would become familiar with the test format and recording process. To this goal, a practice test (Mock test) was played and the researcher recorded their answers. Answers to the practice questions were recorded to confirm that the speaking performance procedure was successful; yet, they were not scored. Also, all the students were given the instruction guide so that they would be able to find out what they were expected to do in details. Following the practice question step, the tasks of both versions of the test were run one by one. Each test was recorded on a video tape for future assessment analysis. Each test was recorded on a video tape for future assessment analysis. The students were given 60 seconds (in all of the semi-direct and four out of five direct tasks) to prepare their responses. Each assessment on the tasks lasted approximately around 12 minutes. The responses were recorded in a MP3 format and saved on a CD for the raters to score.

It should be noted that the 200 students were divided randomly into two groups in a way half of them took the direct and half the semi-direct test version and then the roles were reversed. The reason for not having all the participants perform both versions of the test was due to the fact that performance in one version would most certainly affect their performance on the other version through enabling them to get used to the typology of the questions and this would invalidate the findings of the study. As part of the study which involved close observation of students' performances especially under both direct and semi-direct mediated format, immediately, after the completion of all the test tasks, the students were given the questionnaire and were interviewed along regarding their attitudes towards the test. The addition of the interview, subsequent to the students' questionnaire, as a method of qualitative data collection from the students is to ensure the validity of the research. This triangulation method would definitely shed light on some vague parts of the study which sole quantitative studies suffer from. Thus, having submitted the questionnaire to the research coordinator, each student was precisely interviewed by the researcher to add more assurance to the validity of the job and fulfill the requirements of the triangulation study. It is noteworthy to indicate that the interviews were recorded for more exact data analysis.

Design and Analyses

In order to investigate the research questions, the researcher employed a quantitative quasi-experimental design to investigate the raters' development with regard to rating L2 speaking performance (Cohen, Manion, & Morrison, 2007). Quantitative data (i.e., raters' scores based on an analytic scoring rubric) were collected and analyzed using a MANOVA test to investigate students' score differences of various levels of language proficiency groups with respect to their oral scores in each analytical factor.

RESULTS

RQ: Do students with various levels of language proficiency differ in speaking ability and speech production factors?

As it was already indicated, a sample TOEFL iBT test was administered to make sure that the students participating in this study were at various levels of language proficiency. Table 2 displays the descriptive statistics of the scores obtained by the three groups of students.

Table 2.
Descriptive Statistics of the Test-takers' TOEFL Scores

Students' groups	Mean	SD
Intermediate	74.17	3.01
High intermediate	83.38	3.37
Advanced	90.81	2.71

Also, in order to make sure whether there is a significant mean difference among the scores of the students of the three groups, an ANOVA was run. Table 3 demonstrates the ANOVA statistical analysis of the TOEFL scores of the three groups of students.

Table 3.
ANOVA Table for the TOEFL Scores of the Three Groups of Students

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	23424.620	2	11712.310	2197.362	0.000
Within Groups	1583.060	297	5.330		
Total	25007.680	299			

The outcome shows that there is a significant difference regarding students' general language proficiency among the students. However, since the focus of the study is on raters' scoring of students' oral interaction, the sole assessment of students' general language proficiency for the sake of identification of their differences does not seem valid enough. In other words, the TOEFL test, as a test of general language proficiency, does not account for the students' oral performance; therefore, this provides us with little information about the differences among the three groups of students. Consequently, it was decided to analyze the students' oral performance based on speech production analytical factors. Accordingly, the students' oral productions on each task were analyzed based on *fluency*, *lexical complexity*, *structural complexity*, and *accuracy*.

Fluency measure: there are various measures to fluency. Brooks (2009) identified two various measures for fluency. **Speech Rate A** which is defined as *syllables per minute in unpruned speech* and **Speech Rate B** which is define as *syllables per minute in pruned speech*. The term pruned speech refers to a speech production in which repetitions, reformulations, and false starts have been eliminated, whereas for unpruned speech have not. The main advantage of having both measures of speech rater is that it gives us a more comprehensive view of students' fluency via including both the amount of speech and the speech pauses in a limited time (Brooks, 2009). Both speech Rates A and B are calculated through the following formula:

- ***Speech rate A/B*** *= (number of syllabuses / total number of seconds)×60*

Lexical complexity measure: lexical complexity measure tells us to what extent the students have used lexical words in their speech productions and accordingly to what extent it was lexically complex. Robinson (2001) identified two measures for lexical complexity: *Percentage of lexical words* and *Ratio of lexical to functional words*. The same raw data are used for the calculation of both measures; however, it is related to two category words. The following formulas are used for the calculation of percentage of lexical words and ratio of lexical to functional words.

- ***Percentage of lexical words*** *= (number of lexical words / total number of words)×100*
- ***Ratio of lexical to functional words*** *= (number of lexical words / number of functional words)×100*

Structural complexity measure: the basic measure for measuring structural complexity in oral production is C-unit (Kim, 2011). It is measured as the number of **S-nodes** per **C-unit**. *S-note is a term which is refers to tensed or untensed verbs* and *C-unit is defined as "the main clause plus whatever subordinate clauses happen to be attached or embedded within it"* (Hunt, 1966, p. 735). Therefore, structural complexity is calculated through the following formula:

- ***Structural complexity*** *= (number of S-nodes / number of C-units)*

Accuracy measure: There might be some conceptual misunderstanding regarding the terms accuracy and complexity. According to May (2009) accuracy is investigated through various measures of grammatical morphemes, while complexity is measured by amount of subordination. There are three different measures for calculating accuracy of language performance: the *percentage of error free units*, *language target-like use of articles*, and *error free* clauses (Skehan, 1996). Error free clauses are referred to as "a clause in which there is no error in syntax, morphology, or word order. Errors in lexis were counted when a word used was incontrovertibly wrong" (Skehan, 1996, p. 310). A combination of these measures of language accuracy gives us a comprehensive view of the accuracy of language produced by student involving various angles of language elements. The formula of each of these measures of accuracy is given below:

- ***Percentage of error free units*** *= (number of error free C-units / Number of C-units)×100*
- ***Target-like language use of articles*** *= (number of appropriate used articles / number of obligatory contexts + number of inappropriate use of articles)×100*
- ***Error free clauses*** *= (number of self-repairs / number of errors)×100*

For the last formula, the term "self-repair" could be either self-initiated or other initiated repair which according to Moere (2012) both infer a student's awareness of form and can be

interpreted as his/her attempts at being accurate. Besides, he refers to this as noticing a hole in learners own interlanguage which may lead learners to notice the gap by directing their attention to the correct output. Both these forms of self-repair have been said to potentially lead to acquisition. The outcome of the analysis of all students' oral production throughout the entire study is demonstrated in Table 4 which displays raters' means and standard deviations regarding their function in each language analytical factor for each of the tasks used in the study. Since the maximum score for each task was 7 scores, the mean score of each factor subcategory in each task ranges from 0 to 7 points.

Table 4
Descriptive Statistics of Students Oral Production for Five Oral Tasks

Analysis factor	Factor sub-category	Description						Narration						Summarizing						Role play						Exposition					
		Inter	Inter	Upper	Upper	Ad	Ad	Inter	Inter	Upper	Upper	Ad	Ad	Inter	Inter	Upper	Upper	Ad	Ad	Inter	Inter	Upper	Upper	Ad	Ad	Inter	Inter	Upper	Upper	Ad	Ad
			Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd		Sd
Fluency	Rate A (Unpruned)	4.52	0.04	5.56	0.03	6.78	0.04	3.83	0.03	5.62	0.03	6.15	0.03	3.41	0.05	5.09	0.05	6.12	0.05	4.34	0.05	5.21	0.05	6.51	0.05	4.77	0.06	5.72	0.06	6.35	0.04
	Rate B (Pruned)	4.48	0.07	5.29	0.05	6.62	0.07	3.77	0.02	5.52	0.02	5.86	0.06	3.29	0.03	4.95	0.03	5.94	0.06	4.28	0.05	5.24	0.04	6.48	0.05	4.18	0.01	5.02	0.01	6.27	0.05
Lexical complexity	lexical complexity percentage	4.33	0.04	5.37	0.04	6.54	0.08	4.15	0.02	4.98	0.04	6.23	0.04	3.34	0.03	5.01	0.03	6.01	0.03	4.17	0.06	5.18	0.06	6.36	0.06	4.48	0.05	5.28	0.05	6.52	0.05

	Structural complexity	Accuracy		
lexical to functional percentage	S-nodes per T-units	Error free t-units percentage	TLU of articles	Error-free clauses
4.41	4.18	4.17	3.89	4.08
0.06	0.04	0.04	0.03	0.04
5.29	5.11	5.19	5.03	5.20
0.06	0.03	0.04	0.03	0.04
6.61	6.39	6.22	5.49	6.22
0.06	0.06	0.04	0.08	0.04
3.71	4.22	3.96	3.66	4.11
0.05	0.05	0.02	0.04	0.04
5.45	5.36	5.25	4.99	4.93
0.05	0.04	0.02	0.04	0.04
5.76	6.63	5.94	6.14	6.37
0.05	0.07	0.09	0.04	0.05
3.20	3.12	3.18	3.62	3.05
0.01	0.02	0.04	0.02	0.05
4.84	5.44	4.82	5.14	5.16
0.01	0.01	0.04	0.02	0.05
5.89	568	6.77	6.43	6.17
0.07	0.10	0.04	0.07	0.05
4.29	3.98	4.09	4.16	4.03
0.02	0.08	0.08	0.06	0.07
5.33	4.92	5.09	4.99	4.84
0.02	0.04	0.08	0.03	0.07
6.64	5.92	6.34	6.04	6.35
0.07	0.09	0.08	0.08	0.07
4.28	4.32	4.20	3.82	4.19
0.03	0.04	0.01	0.02	0.01
5.34	5.28	5.24	5.28	5.03
0.05	0.04	0.03	0.02	0.04
6.42	6.41	6.38	5.73	6.69
0.07	0.04	0.03	0.06	0.04

In order to identify whether there exists a significant mean difference among the performance of the three groups of students in each task regarding each of the basics of language analytical factors, a factorial MANOVA was used. Since there were 300 students participating in the study and 8 oral subcategory factors, 2400 data were obtained for data analysis. Table 5 displays the factorial MANOVA results of oral tasks and language analytical factors for the three groups of students.

Table 5.

Factorial MANOVA of Oral Tasks and Language Analytical Factors for the Three Groups of Students

Source	Dependent Variable	Type III Sum of Squares	*df*	Mean Square	F	Sig.
Corrected Model	Description	37479.086[a]	23	1629.525	2.027	.003
	Narration	151507.785[b]	23	6587.295	2.025	.003
	Summarizing	342854.350[c]	23	14906.711	2.028	.003
	Role Play	610478.025[d]	23	26542.523	2.026	.003
	Exposition	955045.873[e]	23	41523.734	2.026	.003
Intercept	Description	644618.704	1	644618.704	801.855	.000
	Narration	708159.615	1	708159.615	217.693	.000
	Summarizing	782756.520	1	782756.520	106.477	.000
	Role Play	854811.015	1	854811.015	65.252	.000
	Exposition	936308.007	1	936308.007	45.678	.000
Students’ levels	Description	11422.226	2	1631.747	2.030	.048
	Narration	47122.125	2	6731.732	2.069	.044
	Summarizing	105327.943	2	15046.849	2.047	.046
	Role Play	187217.865	2	26745.409	2.042	.047
	Exposition	292771.947	2	41824.564	2.040	.047
Analytical factor	Description	3257.108	7	1628.554	2.026	.132
	Narration	13048.208	7	6524.104	2.006	.135
	Summarizing	29690.801	7	14845.400	2.019	.133
	Role Play	52907.520	7	26453.760	2.019	.133
	Exposition	82784.241	7	41392.120	2.019	.133
levels * factors	Description	22799.752	14	1628.554	2.026	.013
	Narration	91337.452	14	6524.104	2.006	.014
	Summarizing	207835.606	14	14845.400	2.019	.013
	Role Play	370352.640	14	26453.760	2.019	.013
	Exposition	579489.686	14	41392.120	2.019	.013
Error	Description	1910089.210	2376	803.910		
	Narration	7729164.600	2376	3253.015		
	Summarizing	17466996.130	2376	7351.429		
	Role Play	31125810.960	2376	13100.089		
	Exposition	48703188.120	2376	20497.975		
Total	Description	2592187.000	2400			
	Narration	8588832.000	2400			
	Summarizing	18592607.000	2400			
	Role Play	32591100.000	2400			
	Exposition	50594542.000	2400			
Corrected Total	Description	1947568.296	2399			
	Narration	7880672.385	2399			
	Summarizing	17809850.480	2399			
	Role Play	31736288.985	2399			
	Exposition	49658233.993	2399			

a. R Squared = .019 (Adjusted R Squared = .010)

b. R Squared = .019 (Adjusted R Squared = .010)

c. R Squared = .019 (Adjusted R Squared = .010)

d. R Squared = .019 (Adjusted R Squared = .010)

e. R Squared = .019 (Adjusted R Squared = .010)

The outcome of the table demonstrates that there exists a significant difference among the performance of the three groups of students from each other (third row). This shows that the students, regardless of what subcategory factor is being considered, differed significantly from each other, $p<0.05$. However, considering the eight subcategory factors, there observed no significant difference among the students of in whole (fourth row). This shows that the students, regardless of their proficiency different levels, did not differ from each other. Nevertheless, when considering both factors of students' various levels of proficiency and the eight different subcategory factors, there observed significant difference $p<0.05$ showing that the students of each level of proficiency performed differently from the other groups (fifth row) in each task with respect to the analytical factors of fluency, lexical complexity, structural complexity, and accuracy of oral language produced by the students.

DISCUSSIONS AND CONCLUSION

The outcome of the research question indicated that students, regardless of what subcategory is being considered, differed significantly from each other. However, considering the eight subcategory factors (see Table 4), there observed no significant difference among the students regardless of their language proficiency levels. Nevertheless, when regarding students' language ability and the eight subcategory factors, once again a significant difference was observed among the test takes. This indicates that students, in each level of language proficiency, were different from each other with respect to various measures of fluency, lexical complexity, structural complexity and accuracy when performing the five oral tasks. Such finding is in line with that of Skehan (1998) and Skehan and Foster (1999) who found that students represented different performances in different task conditions based on different task difficulty factors.

This study also investigated the relationship between students' oral performance abilities and their perceptions of task difficulty measures. The outcome showed a significant relationship between students' perceptions of task difficulty and their oral performance abilities in the dimensions of language planning (when they are provided with preparation time compared to when they are not), perspective (using the first person point of view as compared to the third person) and immediacy (using here-now and compared to there-and-then). This finding is fairly consistent with that of In'nami and Koizumi (2016) who found close relationship between planning time and students' performance. This indicates that students found it easier to speak when they are provided with planning time regardless of their proficiency level. In this respect one of the students commented that "With more planning time, it is possible to better organize your thoughts and put them in speech form and make connections between them." *(A student)*.

The findings also reflected relatively significant relationship between raters' perceptions and the actual task difficulty measures with respect to task dimensions of language planning, perspective and immediacy. Although for the other dimensions some positive correlation was still found, it was not significant enough to be regarded as a determining factor, according to Cohen's table of effect size. On the other hand, rather conflicting findings were found between this finding and that of Skehan and Foster (1999) who found relatively no significant relationship between test-takers' perceptions and their real test performance which might be attributed to contextual differences and test-takers' individual differences.

The above-mentioned finding is relatively parallel with that of Elder, Iwashita, and McNamara (2002) except for planning time where they found no significant difference regarding the manipulation of time preparation on students' oral performance scores. Kim (2011) in a fairly similar study found a relationship between raters' perceptions were task difficulty dimensions; however, his study was based on qualitative data. This finding reflects a connection between task difficulty and students' anxiety. This is relatively in line with what Kuiken and Vedder (2014) and May (2009) who found relationship between students' test anxiety and group oral task and argued that students with low ability suffer from more test anxiety than the ones with higher ability. However, the findings primarily reiterated that the students' oral ability is a more determining factor influencing their scores than their anxiety level. This outcome is parallel with that of Kim (2011) and Davis (2016) who discovered that students' ability but not anxiety is a more important significant variable affecting their scores awarded by the raters.

Rater effects can threaten the validity of decisions made for ratings. The findings of this study demonstrated the usefulness of the use of analytical approaches to rater training programs in detecting rater effects and demonstrating the consistency and variability in rater behavior aiming to evaluate the quality of rating. This study showed that rating oral proficiency tasks is context-specific. The analysis confirmed that the nature of second language oral construct is not constant, thus different results are achieved using different oral task dimensions. Consequently, the outcomes have constructive implications in terms of the use of students' feedback as a reliable indicator of task difficulty factors and specifically as a basis for test design and test validation.

However, the findings must not be misinterpreted as a key factor of establishing a hierarchical order of task difficulty solely on the basis of students' testing intuition. Besides, generalizations must be done with great caution. It is important for performance assessment test to take into consideration the effect of task characteristics and most importantly, performance conditions in estimating the performance ability of students. This research got benefit from five oral tasks in the direct and indirect version respectively. The replication of the research adopting the use of other types of oral tasks could be done in future studies. Besides, since this study considered six factors which were hypothesized to be influential in task difficulty dimension measures, further studies could be run investigating other task testing dimensions on their possible effectiveness of task difficulty.

Finally, care must be done in generalizing the findings of this study since firstly, these were by no means all and only influential factors in task difficulty and for sure there are a number of other dimensions functioning accordingly. Secondly, some of the difficulty factors which were not found as significant correlation between students' perceptions and their actual test performance might be due to the fact that they did not notice them in their test performance thus did not reflect those factors in their follow-up interview/questionnaires.

References

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35-59.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115.

Bijani, H. (2010). Raters' perception and expertise in evaluating second language compositions. *The Journal of Applied Linguistics, 3*(2), 69-89.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

ETS. (2001). *ETS Oral Proficiency Testing Manual*. Princeton, NJ: Educational Testing Service.

Huei-Chun, T. (2007). *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. ERIC Document ED496075.

Hunt, K. W. (1966). Recent measures in syntactic development. *Elementary English, 43*(4), 732-739.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*(3), 341-366.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*(1), 23-48.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment.* (Ph.D.), University of Columbia.

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing, 31*(3), 279-284.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing, 31*(2), 117-204.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499.

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

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 40-56.

Moere, A. V. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344.

Nakatsuhara, F. (2011). Effect of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483-508.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair task performance. *Language Testing, 19*(3), 277-295.

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics, 21*(1), 27-57.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120.

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35-51.

Biodata

D Houman Bijani is an assistant professor in Applied Linguistics, teaching English as a foreign language (TEFL) at Islamic Azad University, Zanjan Branch, Iran. He got his PhD in TEFL from Tehran Islamic Azad University, Science and Research Branch and his MA in TEFL from Allameh Tabatabai University as a top student. He is also an English language teacher and supervisor at Iran Language Institute (ILI). He is a CELTA holder awarded by LTTB Center in Brussels, Belgium. He has published several research papers in scholarly national and international language teaching and assessment journals.

Email: houman.bijani@gmail.com

Dr Adnan Satariyan is a researcher in the field of English as a second language Education at the University of Tasmania in Australia. He received his bachelor's degree in English translation studies and his master's degree in TEFL (Teaching English as a Foreign Language). His research interests centre on pedagogy of literacy skills, metalinguistic awareness, teacher development and leadership, action research, course curriculum, and classroom-based studies.

Email: Adnan.Satariyan@utas.edu.au