International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 3 No. 5; September 2014

Assessing Yemeni EFL learners' Oral skills via the Conceptualization of Target Language Use Domain: A Testing Framework

Sami A. Al-wossabi
English Department
Faculty of Arts and Humanities
Jazan University, Jazan, Saudi Arabia
E-mail: sami_ed@hotmail.com

Received: 13-03-2014 Accepted: 29-04-2014 Published: 01-09-2014
doi:10.7575/aiac.ijalel.v.3n.5p.57 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.3n.5p.57

Abstract

There is an evident lack of a comprehensive evaluation basis for Yemeni learners' speaking skills in the English department, Hodeidah University. The present paper presents a detailed framework of oral assessment criteria that involves a description of target language use domains and then shows how such domains can be systematically related to test design. The framework takes as its main goal the development and description of a criterion referenced rating scale representing real-world criterion elements. The aim of the testing framework, therefore, is to ensure maximum appropriateness of score test interpretations and maximize the validity and fairness of local speaking tests. A five-point likert scale is carried out to elicit 10 trained raters' perceptions of using the pilot scale. The research findings support the use and appropriateness of the scale as it aids raters identify underlying aspects of their learners' oral discourse that cannot be observed in traditional discrete point tests.

Keywords: Target language use domain (TLU), performance based-tests, real language use, rating sale, test fairness, construct validity, language descriptors

1. Introduction

In the current teaching situation, teachers used to rate students usually based on intuition rather than disciplined testing scales whether the speaking task in hand is an interview, a role-play or a presentation. Segmental features of pronunciation and the use of appropriate vocabulary are the only aspects of oral discourse to be evaluated by local raters. However, suprasegmental features of pronunciation, morphological and syntactical features of words and sentences are not attended or even listed in the examiners' score sheet if any. Learners' oral discourse is sometimes recorded and raters subjectively arrive at a score mostly basing their overall assessment on whether learners' speech is intelligible or unintelligible. Hence, one central design decision of the testing framework relates to providing proper rating information to rating scale users. The rating scale, developed as an integral component of the present testing framework, comprises multiple descriptors from multiple sources of information principally associated with linguistic and stylistic features of real language use. The aim is to assist teachers to accurately rate how well a student can speak the language according to pre-defined criteria of different levels of performance. EFL teachers, therefore, can base decisions about test takers' actual performance on multiple sources of authentic discourse-based information, not on traditional constructed-response items and thus tailor effective instructions that fit the learners' needs in their subsequent learning.

2. Background

2.1 Testing speaking skills

Speaking is a difficult task to teach and evaluate particularly in an EFL context where learners have limited L2 environment and teachers use L2 materials that mainly adopt the written norms of language towards more accuracy and at the expense of fluency. Bailey and Savage (1994) stated that "Speaking in a second or foreign language has often been viewed as the most demanding of the four skills …yet for many people, speaking is seen as the central skill" (p. vi—vii). Speaking is, also, as Golebiowska (1990) put it out, "the major and one of the most difficult task confronting any teacher of languages" (p. ix).

Several studies in testing language performance have pinpointed crucial considerations and guidelines for developing and conducting speaking tests. For instance, Gorsuch (2001) hinted at the need for appropriate selection of published speaking tests, as many textbook-based speaking tests do not provide adequate opportunities for learners to exhibit their speaking abilities. Fulcher and Márquez (2003) claimed that one way to reduce task difficulty is to consider cross-cultural differences in developing a speaking test. With regard to Non-verbal language, Jenkins and Parra (2003) suggested that non-verbal strategies should be evaluated in any oral interviews as they can establish an interactive involvement in the same way verbal strategies do. Another consideration that is rarely assessed in oral speaking test is

the evaluation of sociolinguistic rules of speaking. Lazaraton, (1992) significantly pointed out that aspects of conversational interaction such as turn taking, minimal responses and fillers should be developed through effective criteria as testing instruments in assessing any oral proficiency test. Orr (2002) hinted at the necessity to train raters and to encourage them to follow the criteria on which the rating scales are based. The influence of gender on test takers' performance in oral interviews has been also a controversial issue in SLA. O'Loughlin (2002) claimed that gender does not affect the individuals' performance on speaking tests regardless of the gender of raters or the test takers. In contrast, O'Sullivan (2000) claimed that the test takers performed better when interviewed by women regardless of the gender of the participants, as women usually tend to show emphatic support more than men do.

2.2 Performance rating scales

According to Underhill (1987) the purpose of using rating scale is,

> *To describe briefly what the typical learner at each level can do, so that it is easier for the assessor to decide what level or score to give each learner in a test. The rating scale therefore offers the assessor a series of prepared descriptions, and she then picks the one which best fits each learner (p.98).*

This seems to be an overly simplified view of rating scales given the complexity of validity and reliability issues in assessing language performance. Second language inquiry represents a broader scope in second language assessment with multiple perspectives and a wider application of sophisticated testing methodologies (Bachman & Savignon, 1986; Bachman, Lynch & Mason, 1995; Douglas & Selinker, 1993; Fulcher, 1996, 2011; Elder, Iwashita & McNamara, 2002; Matthews 1990; Robinson, 2001).

The literature is also rife with discussions and overviews regarding validity, reliability and appropriateness of the use of performance rating scales (Bachman & Savignon, 1986; Fulcher, 1987; Fulcher & Márquez, 2003; Matthews, 1990; Upshur & Turner, 1995, 1999). Those studies pointed out reliability problems associated with published rating scales, such as, raters' inconsistency, inadequacy of such scales to measure learners' progress in later stages of their developmental learning processes, and usability of rating scales in different learning settings. Validity problems were even more scrutinizingly examined. Those involve the mismatch between scale's descriptors and language features addressed by course objectives, inability of language learners to address some pre-defined descriptors, and ordering of linguistic features in rating scales.

Further, Brindley (1998) argued that a valid rating scale should leave some gap for personal judgments by raters particularly when they are faced with vague descriptors. Nevertheless, raters' judgments could be problematic as it can affect the whole process of performance assessment (Brown, 1995; Caban, 2003; Kim, 2009). Upshur & Turner (1995) argued that two raters of the same student's performance would have different results as each rater has its own interpretation of scale descriptors. Lumley (2005) argued that the raters' agreement on the interpretation of test scores is not because of the rating scale, but is rather "derived from the broadly common experience shared by raters, that of language teaching" (p. 301).

Clearly established criteria for rating of performances can be seen in the study of (Norris, Brown, Hudson, and Yoshioka, 1998, pp. 10) who claimed that performance rating scales should be based on appropriate:

a. Categories of language learning and development

b. Appropriate breadth of information regarding learner performance abilities

c. Standards that are both authentic and clear to students

4. To enhance the reliability and validity of decisions as well as accountability, performance assessments should be combined with other methods for gathering information (for instance, self-assessments, portfolios, conferences, classroom behaviors, and so forth.

2.3 Test Fairness

For the last two decades, the process of test validation has been a central issue in great deal of recent research focusing on the development and use of educational tests. One of the important considerations in using a test is that test must be fair to all candidates and that measures of any test should not weight any bias (Winke, Gass, & Myford, 2013). Such validity, as pointed out by (Roever, 2005), provides optimal opportunities for test takers to exhibit their potential language abilities relevant to the purpose of the test. A test, then, should not exclude test takers on any basis other than the examinee's lack of knowledge. Test takers should be able to present skills the test is intended to measure regardless of age, gender, disability, race, ethnicity, or any other personal characteristics.

Many SLA studies have argued that training raters is a key to increase test fairness, validity of oral assessment and the accuracy of reporting test scores, particularly, when the assessment criteria involves multiple descriptors (Kim, 2009; Elder, Barkhuizen, Knoch, & von Randow, 2007). Fairness is, then, not an isolated concept, but must be conceptualized as an essential element throughout the process of designing and using oral assessment tests. Fairness, for instance, should extend to the accurate reporting of individual and group test results.

The present testing framework has been developed bearing in mind the above mentioned concerns with an aim to provide EFL teachers with well-articulated testing practices that guarantee that teachers operate a fairer testing scale

when interpreting test scores. As Taylor (2006) pointed out, “teaching and testing depend heavily upon having well-described models of language use” (p. 58). Hence, the study provides a description of the target language use domain and the test task, description of tables of specifications, test procedures, scoring method, and description of the scale's descriptors

3. Testing framework

3.1 Rationale

The uniqueness of the proposed framework lies in the fact that it seeks to establish a reciprocal correspondence between real-life tasks and the definitions of actual abilities to be assessed. Such relationship can be seen clearly in the detailed specification of test procedures used to predict inferences of real language use. Further, the framework considers assessment of speech styles that are rarely mentioned in published speaking tests. Speech styles are included in this testing framework as they are important means to initiate any conversational interaction between interlocutors and will show the degree of involvement of students while performing the role-play activities. In addition, the present framework is designed with the EFL Yemeni teaching and learning context in the mind, considering factors, such as, large EFL classes, newness of the proposed testing criteria to local teachers and test takers, and course objectives.

3.2 Description of the test task

3.2.1 Purpose

The test task is designed to provide evidence of students’ ability to converse appropriately in a small interactive talk by role-playing the act of *"Buying Transportation Tickets"* (plane ticket/train ticket/bus ticket). In Addition, the test is meant to provide students with meaningful feedback in order to guide them in their subsequent developmental processes in speaking.

3.2.2 Test type and scoring method

In the same vein, the interpretation of the test scores is based on a *criterion-referenced scale* of multiple descriptors of real language use in order to better describe students according to their potential ability to perform the task in hand. The test type constitutes a part of an achievement test. Students in pairs will role play a task taken from the general content area "encounter services" and the thematic subdivision "choosing among different types of transportation tickets".

3.2.3 Target language use domain

The description of target language use domain (TLU) is adapted from Bachman and Palmer’s model (1996) of TLU task characteristics. The TLU situation is defined as, “a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize” (Bachman & Palmer, 1996, p. 44). The sample of TLU domain for the present project represents three different aspects of buying transportation ticket situations (plane ticket/train ticket/bus ticket).

3.2.4 Construct definition

The construct definition in this test, following a construct -based performance assessment (Bachman, 2002; Norris et al., 1998), is realized via predictions of the test-takers' abilities to accomplish a role-play task. Hence, construct validity is defined as the ability to converse in a small interactive talk in different situations of buying transportation tickets through role-play activities. This ability requires correct syntax, comprehensible pronunciation, adequate and appropriate use of vocabulary and appropriate register. It also requires students to use grammatical, textual, functional and strategic competences by asking/answering questions about transportation tickets (price, class, schedule, time, stops) giving opinions (expensive/cheap prices), etc. Conversation characteristics (speech styles) such as the use of backchannels, fillers will be assessed. The sociolinguistic rules such as register are also assessed. Writing and reading are not tested. Listening is included in the performance but not tested.

3.3 Test design

3.3.1 Description of the test task

The task chosen is a representative sample of the above mentioned TLU domain. It will bear similar characteristics to that of the TLU domain use. The test task is a role-play. The students have the choice to choose among three situations in buying transportation tickets. The targeted students are second year teacher trainees, majoring in English at the English Department, Faculty of Education, Hodeidah University, Yemen. There are 60 students aged between 18 and 22. In pairs, each student can choose with another partner to role-play only one situation, that is, for example, a dialogue about buying plane tickets. In each pair, one student plays the role of a clerk and the other plays the role of a customer.

With regard to the physical characteristics, the location is in a small room in the English department, Faculty of Education, Hodeidah, Yemen. The physical condition is quiet at the time of the activity, well lit, non-distracting. Test takers have the option to bring materials such as maps or schedules. Test takers are familiar with the rater (their teacher) and role-play activities. The participants are the test takers who will play the role of customers, employees, clerk, etc. Each two students should make up their dialogue and act the role-play activity in front of their teacher. The teacher will not take part in the role-play activities.

Considering the characteristics of the test rubric, instructions will be given one week before undertaking the test so that students will have the chance to prepare themselves for the test task. The rubric is written in the target language

(English) in the written channel. Specifications of procedures and tasks are explicit. The structure of the task contains three role-play tasks that involve buying transportation tickets. Five minutes to ten minutes are allotted for each task.

The criteria for correctness are criterion referenced. Students are evaluated on a language ability scale from 1-4 for use of appropriate pronunciation, vocabulary, morphosyntax, and speech styles. Regarding the procedures of the scoring method, only one rater will rate students' performance on a criterion-referenced scale (1-4). The rater will follow pre-defined criteria for rating students.

The language characteristics involve organizational and pragmatic characteristics. The organizational characteristics include grammatical characteristics that are involved in producing accurate utterances using the knowledge of morphology, syntax and phonology. The language domain contains general, formal/informal and frequent vocabulary used in buying transportation ticket situations (tickets, train, plane, bus, etc.). Morphology and syntax consist of primarily organized structures. Phonology represents standard use of speech sounds. Nevertheless, some situations may involve examples of non-standard use of morphology, syntax and phonology. The organizational characteristics also include textual, cohesive and rhetorical characteristics. In the above mentioned TLU domain, cohesion involves the use of a narrow range of cohesive devices, such as pronouns, linking words, adverbs, etc. The rhetorical organization involves clear organizational development of information of language in use.

The pragmatic knowledge involves functional and sociolinguistic characteristics. The functional characteristics in the TLU domain involve ideational and manipulative functions, including requesting, asking for information, accepting, refusing, interrupting, etc. The sociolinguistics characteristics include features such as standard dialect, formal register, natural delivery of language, and minimal cultural references. The topical characteristics are relevant to the type of information and language features that are used in the above situations (e.g. the ability to ask about the direction of flights or the ability to provide information about the price of tickets).

An important category to be considered is seen in the relationship between the input and the response which is defined by Bachman and Palmer (1996) as "the extent to which the input or the response affects subsequent input and responses" (p.55). Such relationship is reciprocal in terms of reactivity. That is, the participants usually exhibit interactive involvement when performing the task in hand. The scope of the relationship is narrow as the relationship between the interlocutors in the above situations is often distant. The directness of the relationship between the interlocutors is direct as responses address specific questions in the input. (See appendix A for a description of the target language use domain).

3.3.1 Description of the table of specification

Many researches claimed that that a better specification of scoring criteria might increase rater's reliability (Hamp-Lyons, 1991; North, 1995, 2003; North & Schneider, 1998). In this testing situation, the table of specifications contains four tables specifying the functions to be assessed with regard to language construct of the present testing framework. As (Chalhoub-Deville, 2001, p. 225) put it out, "Language testers and researchers need to expand their test specifications to include the knowledge and skills that underlie the language construct".

The assessment criteria, therefore, contain language linguistic and stylistic aspects (pronunciation, morphosyntax, vocabulary and speech styles). Organizational features (textual and grammatical organization) are embedded in the description of morphosyntax aspects.

Pragmatic features (functional, sociolinguistic and topical characteristics) are embedded in the description of speech styles aspects. Grammatical and pragmatic features of the TLU are also realized in aspects of task completion (greeting, asking for/offering help, requesting information, providing information, and thanking). Table (1) specifies the total score (100%) that will be devoted in half to linguistic and stylistic aspects (50%) and task completion (50%)

Table 1. Total Score

	Score
Criteria	50
Task completion	50
Total	100

Table (2) presents the criteria for measuring linguistic and stylistic features. In the first column, there are four levels of linguistic and stylistic features of spoken language that will be measured. Each level will be assessed on a scale from 1-4 as shown in column two. Column three shows the weight (actual number) given to each level or criterion which should be multiplied to get the score in column four. Pronunciation is given the least score (10% of the total score) as the test takers are familiar with pronunciation aspects, such as, individual speech sounds, stress, intonation, etc. Speech styles (minimal responses, backchannels, fillers, etc.) are given the most score (16% of the total scores) as the test takers have been recently introduced to aspects of speech styles. Morphosyntax and vocabulary are given 24 % of the total scores.

Table 2. Criteria for measuring linguistic and stylistic features

Criteria	Scale	Weight for the Assessment criteria	Score and %
Pronunciation	1-4	**2.5**	10
Morphosyntax	1-4	**3**	12
Vocabulary	1-4	**4**	12
Speech styles	1-4	**3**	16
Total			50

In table (3), five functions for task completion that will be observed during the test takers' conversational interaction are listed in the first column. A specific weight of the assessment criteria is dedicated to each function. Greeting and thanking are given only 10 % of the total score (50%) as they are fixed formulaic expressions and students are supposed to know how and when to use them. Offering and asking for help, though a kind of formulaic speech, are weighted with 10 % of the total score (50%). Requesting information and providing information are weighted with 30% as they will enable the examiner to observe and assess his/her students' extended oral production and also their ability to use different speech styles. The presence of these functions will be weighted with (1) and their absence will be weighted with (0) as shown in the second column (1-0). The third column indicates the weight of the different tasks and the fourth column indicates the score and the percentages of each task.

Table 3. Five functions for task completion

Task	Completion	Weight for the task	Score and %
Greeting	0-1	5	5
Asking for/ Offering help	0-1	10	10
Requesting information	0-1	15	15
Providing Information	0-1	15	15
Thanking	0-1	5	5

Table (4) displays an overall representation of scoring criteria for both linguistic/stylistic features and task competition.

Table 4. Overall representation of scoring criteria

	Pro	MS	Voc	Ss	TC	Weight for Task completion (%)	score
Greeting					1	10	5
Offering/ Asking for help					1	20	10
Requesting information	1-4	1-4	1-4	1-4	1	30	15
Providing information					1	30	15
Thanking					1	10	5
Weight for the assessment criteria (%)	20	24	24	32		100	
Score	10	12	12	16			50
Total							100

As illustrated above, the test task has a composite score of 100 points that are dedicated in half, 50% for linguistic and stylistic features and 50 % for the task completion. The examiner develops a set of instructions in English to guide the

students in accomplishing the test task (See appendix B). Furthermore, this set of instructions provides the students with the assessment criteria of language aspects and task completion (See appendix C).

The scale includes descriptors that represent features of pronunciation such as segmental aspects (e.g. individual sounds) and suprasegmental aspects (e.g. intonation, stress), morphosyntax (e.g. derivational and inflectional morphemes), vocabulary (e.g. nouns, adverbs of time) and also features of speech styles (e.g. fillers, minimal responses) presented in (Lazarton 1992) and (Biber et al., 1992). The test takers have been introduced to these features throughout four spoken English courses (See Appendix E).

Finally, a score sheet for both raters and students is developed in order to facilitate the recording and the reporting of the assessment information (See Appendix C & D). In the above teaching situation, giving this kind of scoring sheet to students is unusual. However, providing students with this scoring sheet will be of great value as it will not only help them focus on important areas of language ability but also will guide them in their subsequent processing of features of real language use.

3.4 Test procedures

3.4.1 Test takers

The test takers in this speaking test are 60 intermediate second year students who are majoring in English in the English Department, Faculty of Education, Hodeidah, Yemen. They are 20 males and 40 females and their ages are between 18 and 22 chosen from Spoken English (course 4), second semester. They are familiar with the role-play activities as theses activities were regularly being introduced to them earlier in their speaking classes.

3.3.2 Administration

The test will be administered at the end of the speaking course. The test will take place in a small room in the English Department. The sixty students will be divided into 30 pairs. The groups will be tested throughout two days consecutively. Each pair in a group is given only 5 to 10 minutes to perform the role-play activity. The test takers should be given the instructions of the test one week before the test.

3.4.3 Soring procedures

The scale used for assessing the test takers' oral performance is an analytical scale. It is a criterion referenced language ability scale, including four aspects of linguistic and stylistic features (pronunciation, morphosyntax, vocabulary and speech styles). The criterion for each aspect is assessed on a four- band scale (1-4: 1 is poor, 4 is excellent). This part of the test constitutes 50% of the total score of the test task (100%). The test also includes a second part for task completion that is weighted with 50%. Therefore, the test task has a total score of 100%. The teacher will not take part in the role-play activities. The teacher will be the main rater in this speaking test. However, the teacher will select a small sample of the students' scores to be rated by another rater as to provide an acceptable consistency of the rating of the test scores. Then, this sample of test scores, which will be rated by another rater, will be correlated with the teacher's rating of the same portion of score. Such criteria for scoring are operationalized to provide an insight into what raters should pay attention to in the process of rating and, thus, contributes towards the validation of rating scales.

3.5 Plan to evaluate test usefulness

3.5.1 Reliability

Consistency will be across situations. That is, the three different situation of buying transportation tickets should be carefully evaluated in terms of the level of difficulty, performance required to accomplish each task, and the clarity of instructions. There should be an intra-rater consistency following the scoring criteria mentioned above. The teacher will select 10% percent sample of the test score to insure inter-rater consistency. A standard error of assessment will be developed in order to reasonably predict the test takers' true score and its relationship with the observed scores.

3.5.2 Construct validity

The content of the test task should reflect the skills that are to be measured and that could be achieved by providing tasks included in the role-plays that involve the test takers in providing evidence of using, for example, appropriate pronunciation, and use of speech styles. The content of the test task that involves the performance of aspects of language ability should be primarily related to the content of the materials that the test takers have been taught in their speaking courses. It should also be authentic, as it should reflect aspects of target language domain use. Thus, the content of the test should be a representative sample of the relevant language skills that students have been introduced to and that reflect the target language domain use.

3.5.3 Impact

Students should receive meaningful feedback as to guide them in their subsequent learning. The teacher will meet the students after the speaking test in order to discuss and talk about their performance. The students should take part in the discussion and describe their own experiences in preparing for the role-play activities. In addition, decision procedures should be applied uniformly to all groups of test takers. Therefore, we can make sure that all students are treated fairly regardless of the individual test takers group membership.

3.5.4 *Practicality*

Due to the number of students, some considerations should be taken into account. First, only five to ten minutes should be allotted to each role-play task. Second, role-play tasks should be administered throughout two days consecutively so that the teacher can carry on the activities without being exhausted and to reduce the practice effect of role-playing the same test activities. Tasks should not be administered during working hours in order to avoid noise and disturbance. They will be administered in the afternoon after closing hours at 3.00 o'clock.

4. Appropriateness and Usability of the pilot scale

4.1 Methods

A five point likert-scale is conducted to determine the raters' beliefs on using the rating scale. It consists of questions upon which the respondents can express either agreement or disagreement attitudes towards the item in question. Each statement is given a numerical score to reflect its degree of attitudinal approval. The likert scale includes 12 items. The items are grouped into two categories. The first category include 6 items that address the possible limitations of the proposed rating scale whereas the other six questions in the second category address the potential advantages of using the rating scale. It is thought that twelve items would give a good picture of raters' perception of the proposed rating scale considering that all raters chosen to participate are M.Ed. holders and are able to clearly express their stand on the use of the pilot scale.

This likert-scale is typically appropriate to be used in the present study as the purpose is to urge test users and developers to operationalize comprehensive rating scales to ensure validity and test fairness while undertaking oral assessment. Such methodology, however, is thought to be of no value if realized via the involvement of students' opinions on the way their oral abilities are judged. This is because students might greatly be lenient in delivering their true perception of such rating scale. Linguistic and stylistic features involved in the oral assessment could be viewed by many students as difficult to cope with and would necessitate them to do extra effort to incorporate such features in their oral discourse regardless of their importance in any speaking context. Hence, there could be a kind of resistance from learners being judged on multiple aspects of oral discourse and as such there is a great probability of turning down the proposal by stake holders and sticking to traditional discrete-point tests.

4.2 Participants

The participants are ten English teachers, 7 females and 3 males. All of them have M.Ed. in language teaching and education. Their teaching experiences in schools and Hodeidah University range between 3 to 8 years. They have been introduced, in a workshop, to concepts of test validity, test fairness, and the different descriptors included in the proposed rating scale (grammatical, linguistic and stylistic features, etc.) and how they can effectively incorporate them while undertaking the oral assessment procedures.

4.3 Current results and Discussion

Specifically for the present study, in the first category (items 1-6), lower means indicate the raters' disagreement to any possible limitations in the speaking rating scale. Therefore, lower means show the positive side of the likert scale. In the same category, higher means indicate the raters' agreement on the presence of clear limitations in the rating scale. Higher means then represent the negative side of the likert scale. In table (5), as shown below, the average means of items, 1,2,4,5 respectively have lower means and as specified above constitute substantial significance. The specified items are concurrent with the usability and usefulness of the rating scale for oral assessment. Interestingly, in the first category, item3 and item6 show higher average means, though not substantially significant, indicating raters' agreement on two issues. Regarding item3, the raters show noticeable tendency towards the need for special training on the use of the new speaking rating scale. Item 6 indicates that most raters have the feeling that the multiple descriptors involved in the scale could be problematic and difficult for students to cope with. This seems to be normal as it is their first time to operationalize such scale in assessing learners' oral skills.

Table 5. Descriptive statistics of 5-point Likert-scale items from item 1 to item 6: (n:10)

Item	Mode	Mean	Standard Deviation
1	1	1.4	0.516
2	2	2.2	1.033
3	4	3.3	1.25
4	1	1.1	0.316
5	2	1.9	0.738
6	4	3.4	0.966

In the second category, higher means indicate the raters' agreement to the usability and usefulness of the rating scale for assessing oral skills. The overall means in the second category (item7 to item12) are between 4.4 and 4.8. Such results, as specified for the second category, indicate substantial significance towards positive attitudes on the use of the pilot

rating scale. The mean of item12, in particular, is substantially significant. It shows evidently the raters' positive attitudes on the necessity to validate such informative speaking rating scale to be officially operationalized in the teaching context in Hodeidah University. The mean of item8 (4.7) is also significantly informative of the fairness and validity of the rating scale as perceived by the raters.

Table 6. Descriptive statistics of 5-point Likert-scale items from item7 to item 12: (n:10)

Item	Mode	Mean	Standard Deviation
7	5	4.5	0.707
8	5	4.7	0.483
9	4.5	4.4	0.699
10	5	4.4	0.699
11	4	4.4	0.516
12	5	4.8	0.422

Upon individual interviews, the raters revealed that the two-hour workshop was not enough to have a good grasp of the scale descriptors and that they had to examine it several times. Nine raters mentioned that the pilot scale guided them to focus more on different aspects of students' oral discourse. Seven raters indicated that the scale was objective and as such help them easily arrive at a score. All raters felt that the elaborated descriptive scale would provide students with specific information about where they did well, and what they need to work on. In general, the pilot speaking scale was perceived by the raters as positive. However, one limitation that could be noticed is that the pilot scale is a complete novelty for raters and that could affect the way they use it in their rating process. A more prolonged use of the rating scale could result even in more tangible evidence towards the efficacy of such scale to be used in local speaking tests.

5. Conclusion

The present study sought to design a testing framework for assessing Yemen learners' oral skills via the development of a rating scale of multiple descriptors representing features of real language in use. The testing framework, in the present study, underpins the use of a performance-based test approach for oral assessment that is operationalized via the description of the test task in relation to observable domains of target language use. Accordingly, the rating scale is developed bearing features of real world communication. The rating scale, therefore, places primary value upon observations of language performance with an aim to offer the promise of descriptive and complete picture of learners' performance than that of single-criterion rating scales or discrete-point tests. In sum, the present scale is meant to provide meaningful interpretations and inferences from test scores to the type of learners' actual performance in specified domains of target language use.

A vital research goal is, then, to place confidence in the quality of information and interpretation of test scores provided by local raters to the examination board. Another goal is to give guidance for test users and test developers in choosing and selecting appropriate testing tools, delivering valid interpretation of test scores, and providing test takers with appropriate feedback for their subsequent learning.

It is worth mentioning here that a sound and more effective scaling and description of real world language elements that can be traced back to actual performance could be seen in Fulcher's model of Performance Decision Tree (2011). The model is innovative in that it describes pragmatic and discourse variables via a boundary choice approach at arbitrary levels rather than ordering of such variables onto single scale. However, such scale is novel and needs to be put into test to validate its effectiveness for scoring speaking tests in classroom practices.

To conclude, the study's findings support the use and the appropriateness of the rating scale as a measure of speaking proficiency, as well as the utility of the devised discourse-based descriptors for the validation of speaking tasks in other assessment contexts.

References

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476. http://dx.doi.org/10.1191/0265532202lt240oa

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 239-257. http://dx.doi.org/10.1177/026553229901600205

Bachman, L. F. and Savignon. S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *Modern Language Journal 10*(4), 380-90. http://dx.doi.org/10.1111/j.1540-4781.1986.tb05294.x

Bailey, K. M., & Savage, L. (Eds.), (1994). *New ways in teaching speaking*. Alexandria, VA: TESOL.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, England: Pearson Education.

Brindley, G. (1998). Describing language development? In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp.112-140). Cambridge: Cambridge University Press.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.http://dx.doi.org/10.1177/026553229501200101

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), 1-43.

Chalhoub-Deville, M. (2001. Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp.210–28). Harlow, England: Longman.

Douglas, D., & Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing* (pp. 235-56). Alexandria, VA: TESOL.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online rater training program for L2 writing assessment. *Language Testing, 24*(1), 37– 64. http://dx.doi.org/ 10.1177/026553220707151

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. http://dx.doi.org/10.1191/0265532202lt235oa

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28* (1), 5-29. http://dx.doi.org/10.1177/0265532209359514

Fulcher, G., & Márquez Reiter R. (2003). Task difficulty in speaking tests. *Language Testing*, *20*, 321-344. http://dx.doi.org/ 10.1191/0265532203lt259oa

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51. http://dx.doi.org/10.1177/026553229601300103

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria'. *ELT Journal,41*(4), 287-91. http://dx.doi.org/10.1093/elt/41.4.287

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex Publishing Corporation.

Golebiowska, A. (1990). *Getting students to talk*. Great Britain: Cambridge University Press.

Gorsuch, G. (2001).Testing textbook theories and tests: The case of suprasegmentals in a pronunciation textbook. *System, 29*, 119-136.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal, 87*, 90-107. http://dx.doi.org/10.1111/1540- 4781.00180

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral Englishperformance: A mixed methods approach. *Language Testing, 26*(2), 187–217. http://dx.doi.org/10.1177/0265532208101010.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20*, 373-386.

Lumley, T. (2005). *Assessing second language writing. The rater's perspective.* Frankfurth: Peter Lang.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal, 44*(2), 117–121.http://dx.doi.org/10.1093/elt/44.2.117

Norris, J., Brown, J., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report #18). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph 24. Princeton NJ Educational Testing Service

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23(4)*, 445–465.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–263. http://dx.doi.org/10.1177/026553229801500204

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, *19*169-192. http://dx.doi.org/10.1191/0265532202lt226oa

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143-154.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*, 373-386.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 21*(1), 27-57. http://dx.doi.org/10.1093/applin/22.1.27

Roever, C. (2005). "That's not fair!" Fairness, bias, and differential item functioning in language testing. Retrieved from http://www2.hawaii.edu/~roever/brownbag.pdf.

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal, 60* (1), 51-60. http://dx.doi.org/10.1093/elt/cci081

Underhill, N. (1987). *Testing spoken language*. Cambridge: Cambridge University Press.

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. http://dx.doi.org/10.1177/026553229901600105

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49* (1), 3–12. http://dx.doi.org/10.1093/elt/49.1.3

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30* (2), 231-252. . http://dx.doi.org/ 10.1177/0265532212456968

Appendix A

Description of TLU for the Test Task

Test Task **Buying transportation tickets**	
Characteristics of the setting	
Physical characteristics	Location: a small room in the Eng. Dept *Physical conditions:* quiet at the time of the activity, well lit, non-distracting. *Materials:* students may bring their own materials such as a schedule of plane flights. *Familiarity:* Test takers are familiar with each other and with the rater (their teacher).
Participants	*Test takers*: play the role of customers and clerks.
Time of task	Evening, after closing hours: 3.00-6.00
Characteristics of the test rubric	
Instructions	
Language	English
Channel	Written
Specifications of procedures and tasks	Explicit
Structure	three role-play tasks
Time allotment	5-10 minutes for each task
Scoring methods	
Criteria for correctness	*Criterion referenced*: students will be evaluated on language ability scale from 1-4 for range and accuracy for the use of pronunciation, morphosyntax, speech styles, and vocabulary.
Procedures for scoring the method	Raters rate students' performance on a criterion-referenced scale. Another rater will only rate a 10% percent of the students' performance as a sample.
Explicitness for criteria and procedures	Explicit criteria that should be followed by the rater.
Characteristics of the input	
Format	
Channel	Oral: the test takers' interactive talk
Form	Language: verbal/non-verbal

Language	Target
Length	Short to medium delivery of sentences and chunks of speech
Type	Series of items: limited production response, taking the form of an adjacency pair
Degree of speediness	Unspeeded
Vehicle	Live
Language of the input	
Language characteristics	
Organizational characteristics	
Grammatical	*Vocabulary:* specific vocabulary which are used in buying transportation tickets *Syntax:* organized syntactical structures
Textual	*Cohesion:* textually cohesive using cohesive devices such as, pronouns, and fillers *Organization:* organized delivery of individual sentences
Pragmatic characteristics	
Functional	Ideational, manipulative(instrumental) and strategic competence as learners may spend time thinking when there is a breakdown in communication
Sociolinguistics	*Dialect:* standard *Register:* formal *Naturalness:* natural *Cultural reference:* none
Topical characteristics	Relevant to language features and information used in situations of buying transportation tickets, etc
Characteristics of the expected response	
Format	
Channel	Oral
Form	Language: verbal
Language	Target
Length	Relatively short sentences
Type	Limited response
Degree of speediness	Unspeeded
Language of the input	
Language characteristics	
Organizational characteristics	
Grammatical	*Vocabulary:* specific vocabulary which are used in buying transportation tickets *Syntax:* organized syntactical structures *Phonology:* standard pronunciation of speech sounds
Textual	*Cohesion:* textually cohesive using cohesive devices such as, pronouns, and fillers *Organization:* organized delivery of individual sentences and short oral paragraphs
Relationship between the input and response	
Reactivity	Reciprocal
Scope of relationship	Narrow
Directness of the relationship	Mostly direct

Appendix B

Instructions for students

The following is a rubric for the role-play task of buying transportation tickets. The task would be given to test takers one week earlier to the achievement test.

Instructions

In this assignment, you are asked to work with a partner and make up a short role-play activity. This assignment should not take more than ten minutes in length when role-playing with your partner. Both partners should have equal role to play of the situation that they are going to act. You are recommended to get together with your partner and to practice the dialogue several times before acting in front of your teacher. You are required to perform a role-play with a partner about buying transportation tickets. You may choose on of the following situations:

-Buying a plane ticket

-Buying a train ticket

-Buying a bus ticket

In your role-play activity, you and your partner should be able to ask questions and give answers about price of the tickets, class, time, and stop. You and your partner should also be able to give opinions and provide information about prices, (cheap/expensive), class (economy/ first class) time (suitable/not suitable) or any other options that are suitable to the situations. You should also be able to use formulaic expressions that are used in asking for /offering help, particularly those who are going to play the role of a clerk (e.g. *How can I help you?*) and thanking when appropriate.

You have the option to bring materials such as maps for directions of flights, stops or schedules.

Be sure to use appropriate pronunciation, vocabulary, grammatical structures and appropriate speech styles that you have been introduced to throughout this course. Make sure to use formal speech, openings and closings for each role-play task. Informal use is also acceptable. Make sure to provide adequate information whether you are asking or answering, giving information and expressing your opinions

.

Your participation in this test will be evaluated according to the following criteria:

Pronunciation (1-4): individual sounds, stress, intonation, and rhythm, and being intelligible to others

Morophsyntax (1-4): appropriate use of word order, subject agreement, and deviational and inflectional morphemes (past tense and plural markers), clear delivery of sentences.

Vocabulary (1-4): adequate and appropriate use of words clear meaning without instances of hesitance, a variety of usage of parts of speech such as, nouns, verbs adjectives, and adverbs.

Speech styles (1-4): appropriate use of fillers, minimal responses, proper pauses between turn –taking.

Your participation in this test will be calculated by converting the scores for the criteria to a 100 point scale. The passing grade is 60. You will be given a score sheet report of your performance.

Appendix (C)

Student's score report

Name: *Class:*

Aspects of performance (1-4)

Aspect	Score 1-4	Comments
Pronunciation		
Morphosyntax		
Vocabulary		
Speech styles		

Task completion (1 point per function performed)

Task	Completion
Greeting	
Offering help	
Requesting information	
Providing information	
Thanking	

Appendix (D)

Score Sheet for the Examiners

Student: *Class:*

Examiner:

Aspect	Raw Score	Weight (multiply raw score by …)	Final score
Linguistic, and stylistic aspects (1-4)			
Pronunciation		**2.5**	
Morphosyntax		**3**	
Vocabulary		**4**	
Speech styles		**3**	
Total (50)			
Task completion (0-1)			
Greeting		**5**	
Offering help		**10**	
Requesting information		**15**	
Providing information		**15**	
Thanking		**5**	
Total of task completion (50)			
Total score of the task completion and the linguistic, and stylistic aspects:			

Appendix (E)
Score Sheet for the Examiners (Rating Scale)

Student:
Examiner *Class:*

Aspect	Raw Score	Weight (multiply raw score by …)	Final score
Linguistic, and stylistic aspects (1-4)			
Pronunciation 1. Incomprehensible production of speech sounds, inaccurate instances of word stress, intonations problems, unclear rhythm and shows many instances of unintelligibility to the hearer 2. Noticeable mispronunciation of some speech sounds, shortening of lax vowels, stress problems of dissyllabic words, problems of intonational prominence on questions and answers, problems of falling and raising intonation, unclear rhythm, and problems of unintelligibility to the hearers 3. More comprehensible pronunciation of speech sounds within words and sentences, appropriate pronunciation of lax vowels with instances of lengthening short vowels, clear stress with few problems in intonational prominence within sentences, more appropriate use of falling and raising intonation in questions and answers, and very few instances that cause unintelligibility to the hearers. 4. Clear pronunciation of speech sounds within individual words and sentences, strong grasp of word stress and sentence intonational focus, appropriate use of falling and raising intonation of questions and answers, and to somehow a clear rhythm with no instances of unintelligibility.		**2.5**	
Morphosyntax 1. Inappropriate use of derivational and inflectional morphemes (tense /plural markers), unclear use of grammatical structures and organizational aspects within words and sentences such as tense and word order, and subject-verb agreement. Improper delivery of individual sentences 2. Unclear instances of the use of derivational and inflectional morphemes, many instances of inappropriate use of word order and cohesive markers within sentences, many instances of unclear use of grammatical aspects and structures within words and sentences such as tense, subject-verb agreement. 3. More appropriate morphological aspects such as tense markers and plural markers with very few syntactical problems as in the formation of questions (word order), and clear use of grammatical structures such as subject–verb agreement. 4. Very clear morphological features within words and sentences and syntactical features with a strong mastery of word order and organizational structures , organized delivery of individual sentences		**3**	
Vocabulary 1. Inappropriate use of vocabulary items that represent the proper register, inability to recall the right word, inadequate use of words to represent the topic, unclear meaning, and inappropriate use of parts of speech such verbs, adverbs and adjectives. 2. Inappropriate use of vocabulary but can be understood, hesitance in recalling words, somehow clear meaning with many instances of breakdown in communication, problems with the use of appropriate parts of speech. 3. Very few instances of using inappropriate words, good ability to recall appropriate words, and clear meaning, no breakdown in communication, appropriate use of nouns, verbs, adjectives, and adverbs. 4. A noticeable mastery of using appropriate words, good ability to recall words easily, clear meaning without instances of hesitance or inappropriateness, a variety of usage of parts of speech such as, nouns, verbs adjectives, and adverbs.		**4**	
Speech styles(students are familiar with the use of these speech styles) 1. No use of speech styles such as fillers (OK, well), inability to use minimal responses (yeah, hmm), presence of pragmatic issues (lengthy pauses) between turn-taking, many instances of hesitance, inappropriate paraphrasing, and unnecessary repetition, lack of strategic competence. 2. Ability to use the word (OK) but not the words like (yeah and well), somehow longer pauses between turns, few instances of hesitance and unnecessary repetition. 3. Clearer usage of speech styles such as fillers and minimal responses, few instances of longer pauses between turns , inappropriate paraphrasing, no instances of hesitance, presence of strategic competence 4. A strong mastery of manipulating conversations using appropriate speech styles such as appropriate pauses, fillers, minimal responses, appropriate paraphrasing with clear use of strategic competence.		**3**	
Total (50)			
Task completion (0-1)			
Greeting		**5**	
Asking for/Offering help		**10**	
Requesting information		**15**	
Providing information		**15**	
Thanking		**5**	
Total of task completion (50)			
Total score of the task completion and the linguistic, and stylistic aspects (100)			

Appendix (F)

Questionnaire

Rater's Name:

Items	Strongly Disagree	disagree	undecided	agree	Strongly agree
1. The rating scale is time consuming.					
2. The descriptors are by far beyond the students' oral abilities.					
3. The rating scale necessities special training on the part of raters.					
4. The rating scale includes unnecessary items.					
5. Other items should be included o ensure more validity and test fairness.					
6. It is difficult for students to cope with such scoring criteria.					
7. Its descriptors clearly manifest characteristics of real language use.					
8. It broadens the scope for better interpretation of speaking test scores.					
9. The rating scale is fair and accurate to classify students to different levels of performance.					
10.The rating scale helps learners show their true oral abilities.					
11.The rating scale draws the students and teachers' attention to the different linguistic and stylistic features of spoken discourse in real language use.					
12. It should be officially validated for assessing students' speaking skills in the English Dept., Hodeida University.					