The Interaction Between the Type of Rhetorical Modes & the Assessment of Iranian EFL Oral Proficiency

Ahmad Reza Jafari Dehkordi1 - Hossein Bagheri2

Abstract
The present study aimed to investigate the interaction between four modes of rhetorical discourse and the type of rating procedure used for assessing the speaking ability of Iranian EFL learners. A corpus including the recorded speech of 32 advanced EFL learners was subjected to two methods of rating and also analysed with respect to four rhetorical discourse representations of narration, explanation, description, and argumentation. Four competent raters rated the recorded speech once based on a holistic set of criteria and also according to an

1. Lecturer, Islamic Azad University, Shahrekord Branch
E-mail: arjafari@iaushk.ac.ir
2. Lecturer, Islamic Azad University, Shahreza Branch
analytic basis (ACTFL). The obtained data was then subjected to a two-way ANOVA. The results indicated the differential performances of the candidates, which was argued to be attributable to the very nature of the four rhetorical modes of spoken discourse in that each mode of discourse demands certain linguistic structures and a number of syntactic, semantic, pragmatic, or discoursal formulations.

**Key words:** Rhetorical Modes, Assessment, Speaking Ability, Rating Type, Iranian EFL Oral Proficiency.

1. Introduction

The question of L2 speaking ability in general and its assessment in particular has received relatively scant attention in the realm of TEFL / TESOL literature. Recently, oral proficiency testing has received much attention, and many different types of oral tests have been constructed. However, unlike tests of reading, writing, and listening, most of these oral tests have not been put to large-scale validation studies (Taylor, 2000).

Research into the issues concerning the testing of spoken language dates back to a few decades ago. During the 1980s, a large number of researches were conducted on speaking tests produced and used in North America, emphasizing the issues of content validity, construct validity, concurrent validity, reliability, and rating procedures (Taylor, 2000). In 1990s, the main concern for language testing specialists was to determine different factors interacting in the context of oral proficiency assessment. Many established tests of oral proficiency focus on accuracy and consider spoken language as something which can be divided into separate components (Cambridge Oral Exam, SPEAK [Speaking Proficiency in English Assessment Kit]), The Finnish National Certificate Scale, and The Common European Framework Speaking Scales (CEF).

On the face of it, testing, interpreting the scores, and finally making decisions in L2 speaking tests are associated with some kind
of inherent subjectivity on the part of raters. However, this does not mean that L2 oral proficiency is so elusive a phenomenon which cannot be evaluated objectively. As Underhill (1987) posits, oral exams are likely to be less reliable than other kinds of tests simply because they require a subjective judgment by a person about another. EFL/ESL teachers usually rank their learners in terms of their L2 speaking ability along a ranking scale. They may also assign marks as to their learners’ speaking proficiency based on different marking systems. However, the fundamental issue is the common practice of assessing learners’ speaking ability irrespective of the rhetorical modes of discourse of the topics they speak about. In other words, some students might outperform others; for instance, when talking about a controversial issue through argumentation rather than giving a description about a place or explaining a phenomenon.

Therefore, it seems that by requiring all learners to talk about a single topic in one particular discourse mode, some learners might be at a conspicuous disadvantage due to their background knowledge, experiences, their lack of familiarity with the topic(s), or other personal idiosyncrasies, while some others might receive some privileges simply because of one or a combination of the above reasons. Thus, both the reliability and validity of the administered tests might turn out to be suspect.

Knight (1992) posited a number of assessment criteria for L2 oral proficiency such as grammar, vocabulary, pronunciation, fluency, conversational skills, sociolinguistic skills, nonverbal skills, and content. Bachman (1990) proposed another approach in defining language proficiency called "the interactional/ability approach". In this approach, language proficiency is defined in terms of its components abilities such as those described in the framework of Halliday (1976 cited in Bachman, 1990), or the communicative frameworks of Munby (1978 cited in Bachman, 1990) and Canale and Swain (1980).

Research involving oral language tests has begun to address the
discourse produced in these interviews. In a study Shohamy et al. (1998) maintained that many topic-based tests of language proficiency used in university admission procedures (CAEL, IELTS) claim to have pedagogical validity by reflecting language use in academic domains. One possible threat to this validity, as Shohamy et al. (1998) contend, is test bias in that familiarity with the test topic and background knowledge may influence test-taker performance. Fulcher and Márquez Reiter (2003), in an article reviewed assumptions underlying approaches to research into speaking task difficulty and questioned the view that test scores always vary with task conditions or discourse variations.

Luoma (2004) proposed a modular approach to construct, task, and assessment specifications. Her modular specification divides the task of specification writing into focused chunks. Taguchi (2007) investigated the impact of task difficulty on oral speech act production. He found that there was a significant proficiency effect on appropriateness ratings and speech rate, but not on planning time. Galaczi (2009), advocated the use of both qualitative and quantitative methodologies in the design of performance assessment scales for speaking.

In holistic rating the rater awards a score on the basis of a learner’s overall performance, without referring to any special characteristics or using counting errors within a particular system (Underhill, 1987). This type of assessment could also be made with reference to a rating scale, on a single scale for the whole test or for certain individual mark categories. Larry Davis (2009) posits, in his study, that a test taker brings a language proficiency identity (LPID) to a test task. With the advent of a shift in emphasis to language as a means for communication, the more modern style of mark categories requires a consideration of the speaker and the context of speaking.

As for rhetorical modes, McCarthy (1991), found that spoken rhetorical discourse modes can be analysed for their typical patterns.
and linguistic realisations that accompany them. (E.g., service encounters, telephone calls, lectures, etc.). Van Dijk (1989) believes that “if discourse types were merely differentiated based on their different semantics content (topic), we would have a potentially infinite number of discourse types”. He says that we need a more abstract theory, which includes rules of a more pragmatic kind, relating to the specific functions of discourse types in certain contexts.

This study aims at revisiting the existing criteria in L2 speaking tests as well as foreign language learners’ performance on a speaking test with regard to rhetorical modes of discourse. In addition, two conventional methods of L2 oral proficiency rating, namely ‘holistic rating’ as opposed to ‘analytic rating’ will be addressed in the current research.

2. Method
2.1. Participants
Thirty-two graduate EFL students (16 male and 16 female) participated in this study. These students studied at Isfahan University, Tehran University of Science and Technology, and Islamic Azad Universities; Khorasgan and Najafabad branches. It seems that regarding the validity and reliability of English oral proficiency assessment, graduate EFL students can be far more eligible candidates in that they can show their speaking command in different contexts more readily and effectively. Among the total number of 150 EFL learners in all of the above universities, 90 students were selected randomly. Out of these 90 students, 60 students with higher GPAs were selected to participate in a general English proficiency exam (TOEFL) as well as Cambridge Examination of Spoken Discourse. In the final stage, 32 subjects were selected based on the scores they obtained on a standard general proficiency test in English as well as their scores on the oral proficiency test. Attempts were made to choose eight homogenous subjects from each university.
2.2. Instrumentation
Three different tests were administered to the participants in this study. The first test was an English general proficiency test (TOEFL Test) which is an instrument to gauge examinees’ overall proficiency in English. Regarding the fact that TOEFL test per se cannot be a reliable and valid indicator of the participants’ speaking ability in this research, the second test—a standardised speaking test of English (Cambridge Examination of Spoken English)—was administered to the participants. The third test was a speaking test of English comprising four different topics with embedded narrative, descriptive, explanatory, or argumentative rhetorical modes.

2.3. Procedure
To secure the homogeneity of the participants, the selected 60 EFL learners were given the TOEFL Test. This was intended to gain measures of the subjects’ overall performance in English. After the selection of the homogeneous subjects on the basis of their scores on the test of "general English proficiency”, the participants were given a standardised speaking test of English (Cambridge Examination of Spoken English) which did not have any specific orientation in terms of rhetorical modes. Cambridge Examination of Spoken English is one of the oral proficiency measuring devices the reliability and validity of which have already been established by UCELS.

In the third stage, the specifically “discourse mode-oriented” speaking test of English was administered to the participants. It is worth noting that constructing such a test is a highly meticulous and at the same time a demanding task. Despite these apparent difficulties, caution was exercised to construct a test which, relatively speaking, conforms to the proficiency guidelines and norms elaborated by the ACTFL. Also attempts were made to create a friendly and natural atmosphere for the oral exam in order to obtain more reliable results. To that end, the researcher first presented oral instructions concerning
the topics so that the participants could get a nodding acquaintance with what they were supposed to talk about. In the final phase of preparation the respective topics were shown to the participants in written form to assure their awareness of different topics each of which entails a different rhetorical mode of discourse. It must be noted that in order to take care of “Observer’s Paradox” to some extent, the participants were not aware that their voices were being recorded.

3. Results

A two-way ANOVA was run within the SPSS Program to verify the possible impact of the independent variables statistically. The results of the Analysis Of Variance (ANOVA) are depicted in “Table 1” below.

Table 1: Results of the Two-way ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
<th>Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Mode</td>
<td>128.187a</td>
<td>4</td>
<td>32.047</td>
<td>55.485</td>
<td>.000</td>
<td>.469</td>
</tr>
<tr>
<td>Intercept</td>
<td>7293.808</td>
<td>1</td>
<td>7293.808</td>
<td>9196.045</td>
<td>.000</td>
<td>.997</td>
</tr>
<tr>
<td>Scoring Method</td>
<td>59.098</td>
<td>1</td>
<td>59.098</td>
<td>102.319</td>
<td>.000</td>
<td>.290</td>
</tr>
<tr>
<td>Rhetorical Discourse Mode</td>
<td>69.090</td>
<td>3</td>
<td>23.030</td>
<td>39.873</td>
<td>.000</td>
<td>.323</td>
</tr>
<tr>
<td>Error</td>
<td>144.973</td>
<td>251</td>
<td>.578</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7566.968</td>
<td>256</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>273.160</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared=.469 (Adjusted R Squared=.461)
The figures in Table 2 below indicate the results of the comparison between the two ‘methods of scoring’ employed in this research.

**Table 2: Pairwise comparisons of Holistic vs. Analytic Scoring Methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Difference</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>95% Confidence Interval for Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>Analytic</td>
<td>.961</td>
<td>.095</td>
<td>.000 - 1.148</td>
</tr>
<tr>
<td>Analytic</td>
<td>Holistic</td>
<td>-.961</td>
<td>.095</td>
<td>-.000 - -.148</td>
</tr>
</tbody>
</table>

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

In accordance with the figures in Table 2 above, the mean difference for the two “methods of scoring” employed in this study is statistically significant at the .05 level. This could mean that ‘holistic’ rating as opposed to ‘analytic’ rating has rendered different results, thus different assessments by the raters and differential performances on the part of the EFL candidates.

As Hatch and Farhady (1981) put it, researchers most often do comparisons after the analysis due to the fact that they might not have been able to make hypothesis before running the ANOVA. Sheffe test is one of the most common methods of what statisticians call ‘post hoc comparison’. Moreover, the combination of the ANOVA and the Sheffe test makes it possible to find out whether the levels of one independent variable differ in how they affect performance on the dependant variable.
L2 Oral Proficiency

Table 3: Scheffe' Test. The error term is Mean Square (Error)=.578 Alpha=.05

<table>
<thead>
<tr>
<th>Subset</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>64</td>
<td>14.2803</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explanatory</td>
<td>64</td>
<td>14.6813</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argumentative</td>
<td>64</td>
<td>15.2391</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narrative</td>
<td>64</td>
<td></td>
<td>15.6397</td>
<td></td>
</tr>
<tr>
<td>Sig.</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4: Multiple comparisons of four different rhetorical modes of discourse

<table>
<thead>
<tr>
<th></th>
<th>95% Confidence Interval</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>Descriptive</td>
<td>1.3594*</td>
<td>.1343</td>
<td>.000</td>
</tr>
<tr>
<td>Explanatory</td>
<td>.9584*</td>
<td>.1343</td>
<td>.000</td>
</tr>
<tr>
<td>Argumentative</td>
<td>.4006*</td>
<td>.1343</td>
<td>.033</td>
</tr>
<tr>
<td>Narrative</td>
<td>-1.3594*</td>
<td>.1343</td>
<td>.000</td>
</tr>
<tr>
<td>Explanatory</td>
<td>-.4009*</td>
<td>.1343</td>
<td>.033</td>
</tr>
<tr>
<td>Argumentative</td>
<td>-.9588*</td>
<td>.1343</td>
<td>.001</td>
</tr>
<tr>
<td>Narrative</td>
<td>-.9584*</td>
<td>.1343</td>
<td>.000</td>
</tr>
<tr>
<td>Descriptive</td>
<td>.4009*</td>
<td>.1343</td>
<td>.033</td>
</tr>
<tr>
<td>Argumentative</td>
<td>-.5578*</td>
<td>.1343</td>
<td>.001</td>
</tr>
<tr>
<td>Narrative</td>
<td>-.4006*</td>
<td>.1343</td>
<td>.033</td>
</tr>
<tr>
<td>Descriptive</td>
<td>.9588*</td>
<td>.1343</td>
<td>.000</td>
</tr>
<tr>
<td>Explanatory</td>
<td>.5578*</td>
<td>.1343</td>
<td>.001</td>
</tr>
</tbody>
</table>

* The mean difference is significant at the .05 level.
According to the figures in table 4, the participants’ performances in the administered test of oral proficiency show remarkable variation in different rhetorical modes of spoken discourse. In Figure 1 below, the mean difference of the candidates’ scores on each of the four rhetorical modes of discourse as well as the two methods of scoring are depicted. According to Figure 1 below, there exist some conspicuous differences among the four different rhetorical discourse modes in terms of the obtained means. Narrative discourse, relatively speaking, enjoys the highest mean whereas descriptive discourse has been apparently the least manageable mode of discourse for the participants in this study.

![Figure 1: Estimated Means of Four Rhetorical Modes of Spoken Discourse](image-url)
In this study, a corpus of data including the recorded speech of 32 EFL proficient learners was subjected to two methods of rating (Holistic vs. Analytic) on the one hand, and was analysed with respect to four rhetorical discourse modes of narration, explanation, description, and argumentation, on the other. Four experienced and competent raters were asked to rate the speech produced by the candidates twice; once based on a holistic (impression) set of criteria, and in the second stage on an analytic basis (ACTFL Guidelines for Oral Proficiency Assessment.

Drawing upon the descriptive statistics, the following statements can be made:

**First**, the values for \( F_{\text{observed}} \) and \( F_{\text{critical}} \) show that \( F_{\text{observed}} \) for “Scoring Method” is much higher than \( F_{\text{critical}} \). Therefore, this could indicate the significant impact of “Scoring method” on EFL oral proficiency assessment.

**Second**, the \( F_{\text{observed}} \) for “rhetorical Mode of Discourse” is quite greater than the \( F_{\text{critical}} \). In this way, it can be implied that “rhetorical mode of Discourse” has had a significant effect on the differential performance of Iranian EFL learners in a test of speaking. Thus, this shows the conspicuous impact of “rhetorical mode of discourse” on EFL oral proficiency assessment.

**4. Discussion**

Drawing on the obtained results, there were two salient findings in this study. First, The differential performances of the participants with regard to “method of scoring” on the one hand, and second, the participants’ differential performances according to “rhetorical modes of discourse” on the other. Each of the above findings warrants plausible interpretations or comments (Figure 1). As for the candidates differential performances in terms of ‘method of scoring’ it could be maintained that this remarkable variation cannot be regarded as
unexpected or far-fetched due to the fact that in most of the other domains of language performance testing such as writing, for instance, coming up with noticeable differences in terms of the candidates’ performances measured on the basis of ‘holistic’ as opposed to ‘analytic’ approaches and band scales cannot be that uncommon. Therefore, it can be contended that this apparent difference with regard to “method of scoring” in a speaking test is a matter of degree. In other words, here the magnitude of difference is what counts, not simply the existence of this difference as a whole.

Concerning the second observed remarkable variation, that is the possible impact of various ‘rhetorical modes of discourse’, it seems that the picture is a little different. Figure. 1 shows that the estimated mean for ‘narrative’ mode of spoken discourse is the highest among the four rhetorical modes of discourse employed. As put by some researchers in the field, (Labov, 1995; Larsen-Freeman, 1980), narrative discourse is by far the most common type of discourse especially the spoken discourse. However, there ought to be an important premonition: the idea that candidates may have personal idiosyncrasies in term of their ability to narrate something through speech cannot be ruled out. Regarding the three other modes of ‘explanation’, ’description’, and ‘argumentation’, the question of allowing for the candidates personal idiosyncrasies including their background knowledge, amount of prior practice, as well as their possible prior exposure to each of these rhetorical modes of discourse, is of paramount importance. Another interesting finding in this study was the conspicuous variation with regard to the candidates’ performance on the four ‘rhetorical modes of discourse’. The statistics show that ‘narration’ has been, relatively speaking, the easiest or the most manageable ‘mode of discourse’ for the candidates, whereas, ‘description’ has been presumably the most difficult ‘rhetorical mode’ for the same candidates.
This can be accounted for by two possible interpretations. It could be said that this apparent difference in the candidates’ scores on the four different modes of discourse might be due to the nature of these ‘rhetorical discourse modes’. Second, it can be claimed that this difference could be due to the phenomenon that each topic or genre poses particular burden on the speakers. In other words, each mode of discourse demands certain linguistic structures and a number of syntactic, semantic, pragmatic, or discoursal formulations which must be, by nature, more complicated than other modes thus resulting in different level of spoken discourse. To cap it all, it could be said that the apparent differential performances of the candidates in this study could be due to the very nature of these rhetorical modes of spoken discourse candidates dealt with. Alternatively, it could be the result of the interaction of a number of variables which are unaccounted for.

5. Conclusion
The scope for the use of multi-faceted Rasch measurement in research on the performance tests, as put by McNamara (1996), is great because of the complexity of factors which have to do with the interaction of such variables like candidates, raters, and test environment created by the performance requirement. Moreover, we have seemingly a long way ahead to come up with validated rating scales or L2 speaking performance level descriptors-descriptors that play an important role in the definition of the construct validity of performance testing in general. As Louma (2004) maintains, an important issue in assessing L2 speaking is to determine the possible effects of the performance of one interlocutor on the other(s). Furthermore, the pivotal role of task-orientated performance measurement within the realm of oral proficiency assessment is nowadays among the most intriguing areas of additional empirical investigations.
The question of measuring L2 oral proficiency and that of measuring L2 performance in a broader sense have yet to be probed by more exhaustive and empirical and non-empirical researches; researches which help to unveil a small part of the long-standing misconceptions or misinterpretations regarding a great many of the unaccountable issues surrounding the phenomenon of language as a whole.

The results of this study merely shed light on the still-complicated nature of spoken discourse the bulk of which is still in the darkness. The findings can contribute to the construction of criteria and/or modification of the existing norms and standards as well as the establishment of criteria to be used in rating scale construction. They can also help test specialists devise a number of speaking tasks which are specifically geared to the learners’ communicative needs.

References
The Interaction Between the Type of ... / 99

Received on June 11, 2010, from : www.surrey.ac.uk\Eli\Ilta\Ltrc


Shohamy, E., & Fox, J. et al. (1998). Validating a topic-based test of language proficiency. LTRC98. Los Angeles, California. Received on May 25, 2000, from www.surrey.ac.uk\Eli\Ilta\Ltrc


agreement in testing second language oral proficiency. LTRC98. Los Angeles, California. Received March 15, 2000 from www.surrey.ac.uk\Eli\Ilta\Ltrc