ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 8, No. 4, pp. 679-690, July 2017
DOI: http://dx.doi.org/10.17507/jltr.0804.06

A Comparison between Holistic and Analytic Assessment of Speaking

Maha Ounis
University of Sfax, Tunisia

Abstract—**As speaking is an intricate productive skill, its evaluation raises many interesting issues. In this study, the main purpose is to find the most suitable procedure that is useful to assess speaking, by making a comparison between the holistic and the analytic scales. Thus, twenty oral performances were recorded and given to five raters who assessed them holistically and analytically, according to four distinct tasks at the Higher Institute of Languages in Gabès. The raters' gradings were compared according to four basic parameters, namely the mean, the range, inter-rater and intra-rater reliability. Despite the fact that both methods yielded low reliability rates, it was clear that there were many differences between the two scales, since they reflected a discrepancy at the level of the four parameters as well as at the level of tasks. Apart from the differences, it was noted that the holistic scale is more useful, reliable and consistent as far as speaking assessment is concerned.**

Index Terms—**holistic assessment, analytic assessment, speaking**

I. Introduction

Speaking is a skill that reveals a lot about the person's thoughts, personality and origin. Some researchers claim that speaking has its own rules which differ from one person to another and from one culture to another. In addition to that, a fluent speaker makes his audience attracted to his speech, because this shows the extent to which such a speaker masters the language.

For these reasons, many researchers highlight the importance of speaking as part and parcel of the speaker's personality. Hence, if an interlocutor makes many pauses while speaking, the listeners think of his/her lack of confidence or of his/her inability to express himself/herself in a second or foreign language. Consequently, the listeners may judge a speaker according to the way s/he speaks.

In this research, I am interested in speaking in an academic context, particularly "testing speaking". In fact, many issues come out from the association of "testing" with "speaking". Many researchers dealt with testing speaking as a challenging task which faces the testers, for many reasons. Firstly, the momentary nature of speaking urges the testers to listen carefully to the testees' speeches, which is not always guaranteed since there are some internal (tester's fatigue, testee's low voice etc…) and external (noise, the lack of means of recording etc…) factors that impede the assessment process. Secondly, speaking is an intricate skill. Hence, even the definition of this productive skill differs from one researcher to another. In this regard, some researchers consider a good speaker as the one who is able to appropriately convey a given message at a specific context, whereas other researchers take into consideration the different parts of the speech and insist on the necessity to focus not only on the message itself but also on its accuracy. (Hughes (1989), Alderson et al (1995)).

In other words, there are two distinct approaches to the study of speaking. On the one hand, the first view consists of considering speaking as a "whole", so that the tester focuses on the message itself regardless of its parts. On the other hand, the second view reflects that any speech is partitive and can be divided into its subcomponents (like pronunciation, fluency, accuracy, content etc…). In fact, the idea of this research is based on these two conceptions.

II. Hypothesis

This study focuses on a comparison between holistic and analytic scales, in order to determine the most suitable way to assess speaking. This comparison is based on some parameters that enable us to measure the degree of difference between the two methods. Thus, it is hypothesized that the holistic and analytic methods of assessment are different in terms of mean, range, inter-rater and intra-rater reliability and that the variability of speaking tasks deepens this difference, as a starting point to tackle further issues related to assessing speaking.

III. Antecedent Research

If we look at "speaking" from a psycholinguistic angle, we can imagine a scale with two extremes, ranging from "babbling" to "speaking". For instance, some psycholinguists are interested in studying the evolution of a person's oral output since his/her birth. In this context, Caldwell (2009) suggests that the child's speaking progression lies in "producing sounds that will eventually be shaped into words and used in communication." (p.1).

The writer further maintains that even deaf children babble, which makes this primary form of output universal. In fact, the passage from “babbling” to “talking” has a natural aspect: All normal children acquire words and sentences progressively as they grow up. This naturalness, however, does not nullify the complexity of oral acquisition, since many factors play a great role in shaping one’s oral performance. Accordingly, we cannot ignore the role of the environment in enhancing or impeding the child’s speaking progression.

Speaking is also an issue that attracts many discourse analysts, like Fasold (1990) who reports that: “The viewpoint succinctly stated by the famous American linguist Leonard Bloomfield (1933:21) went unchallenged: ‘Writing is not language, but merely a way of recording language by means of visible marks’” (p. 276). Hence, speaking as a skill came before the process of writing.

In addition to that, many researchers deal with speaking in second or foreign language. One of these researchers is Richards (2008) who states that “Research has also thrown considerable light on the complexity of spoken interaction in either first or second language” (p.1). Thus, we can deal with speaking as a complex pattern that necessitates a deep analysis depending on its context. In this regard, Luoma (2004) distinguishes between a planned speech and an unplanned one. A planned speech is the one which is prepared in advance, like a lecture, whereas an unplanned speech is spontaneous like in a conversation. It has to be noted that the planned speech is frequently found in a classroom context, where the teacher can, for example, prepare a lecture, and a student can also prepare an oral presentation.

Assessing speaking is an intricate task. In this context, Luoma (2004) maintains that there is a “cycle of assessing speaking”. She claims that there is a big deal of “interaction” in this type of assessment. Firstly, she specifies the different participants in the stage of testing an oral performance, namely “the examinees, interlocutors, raters and score users”. (p. 4).

Before dealing with the “holistic approach”, it is necessary to know what the word “holism” means. Hornby (2000) defines “holism” as: “Considering a whole thing or being to be more than a collection of parts”. (p. 620). Thus, the term “holism” refers to the whole of anything.

Holistic assessment aims at looking at the overall performance under testing. In this type of assessment, the skill that is tested is considered as a unified whole, where there is no possibility of assessing each part independently of the other. In this regard, Xi et al (2006) maintain that: “In holistic scoring, raters consider the combined impact of delivery, language use, and topic development, and make a judgment about a person’s performance on a particular task. During this process, raters attempt to weigh the impact of different dimensions on the overall effectiveness of communication to come up with a holistic score”. (p 32).

According to this view, in a holistic assessment of speaking, the testers look at the overall oral performance of the testee. The different components, namely fluency, vocabulary, pronunciation etc…, are assessed under speaking as a whole.

The analytic method of assessment is viewed differently by many researchers. In a definition of analytic assessment, Xi and Mollaun (2006) argue that: “An alternative scoring approach, analytic scoring, can be used to assess examinees’ performance on each of the three dimensions. In other words, separate delivery, language use, and topic development scores can be reported”. (p.1). We can guess from this definition that analytic assessment is based on separating the components of the skill under testing. Hence, assessing speaking necessitates providing a score for each subcomponent of speaking, like pronunciation, vocabulary, the message content etc… (Hughes (1989), Alderson et al. (1995), Mertler (2001)).

IV. Research Design

To begin with, I chose that this study be a cross-sectional one, because it aims at looking closely to groups of students and teachers at the same time. Since this research belongs to language testing as a discipline, we cannot imagine a test without taking into account different students and teachers, because a test and its scoring are not actually targeting one student or one teacher.

Despite the fact that this project is mainly based on the qualitative method, the quantitative aspect is crucial since it serves as the first step towards making a comparison between the holistic and the analytic methods of testing. In addition to that, after giving the recording of the students’ oral performances to the raters, it is crucial to draw some tables and diagrams reflecting the statistics that are necessary to make the comparison.

As for the main skeleton of the methodology, it consists of selecting two groups from first and second years English students at the Higher Institute of Languages in Gabés, Tunisia, and five teachers of oral expression from different universities. All the teachers studied Linguistics as a branch, so they are acquainted with language testing as a discipline in general, and testing speaking, in particular. In a further step, I recorded the students’ oral performances according to different speaking tasks, and these performances were then downloaded in a computer by means of a mobile phone. I gave the recorded performances to the five teachers, and then they assessed them holistically and analytically.

After collecting the data, I made a comparison between the holistic and the analytic methods of assessment, according to different criteria that are basically related to language testing namely, inter-rater and intra-rater reliability. In this regard, I used Nakamura’s (2004) comparison between holistic and analytic assessment of writing, by applying his framework to assessing speaking.

A. Audio Recording

Audio recording helps the testers in assessing oral performances, since they can have some reference against which they can check the words, the chunks or even the sentences that are not previously heard. In addition to that, Alderson et al (1995) stress the fact that audio recording offers "authentic language data in tests" (p.224). Accordingly, it positively contributes to any research which is about speaking, because this skill requires authenticity.

Thus, I tried to record some of the students' oral performances, in order to show how the same performance is judged holistically and analytically by some teachers of oral expression. Hence, I used my mobile phone and my computer in order to record and to download the voices. Contacting some teachers and students during and after oral expression sessions was the first step towards recording the speeches of twenty students.

B. The Tasks

In order to ensure the consistency of scoring the oral skill, I gave different tasks to different students. These tasks were actually the ones that some teachers of oral expression gave to their students. As a matter of fact, I tried to attend some oral expression sessions, so that the students were able to think about the major parts of these tasks and then were asked to speak accordingly. All of these performances were later recorded. In this context, the students were assessed on four different tasks namely, oral presentation, interview, storytelling and picture description.

Oral presentation concerned some of the first year students who were asked to deal with this topic "Why have you chosen English as a branch of study?" It has to be noted that the students were given two minutes in order to think about the topic, and then I recorded their speeches. Moreover, I asked them not to take notes before speaking, in order to preserve one essential feature of any speech, namely authenticity.

I have to mention that some of the students were not able to speak at all, so they kept silent, while there were other students who spoke enthusiastically about this topic.

After attending an oral expression session of the first year, I recorded five oral performances in an interview with a teacher. During the session, the teacher asked a question: "What do you think about Valentine's Day?", and she chose one student to ask him/her more questions, and so on. Then, the teacher enabled me to record some interviews.

I noticed that the students were so enthusiastic while speaking in an interview. They were even dealing with the different reasons explaining their attitudes towards Valentine's Day. In fact, I guessed that the teacher wanted to make the students able to speak in a specific context, so that she was able to make them talk about current events.

In storytelling task, the teacher gave a group of second year students a string of words, namely "proper, afraid, foot, seventy-five, quickly", and she asked them to imagine a story using these words. After four minutes of reflection on the task, I recorded some of the students' performances. Each student spoke about his/her own story, despite the fact that there was the same set of words.

In picture description task, I showed some of the second year students a picture, which consists of an anti-war slogan. In this picture there is a child who is expressively pointing to the viewers, with the slogan "Our world needs you to end war!" Some students tried to depict this picture showing their opinions about the child, the colours and the slogan.

C. The Questionnaire

It is essential to deal with the quantitative aspect which enabled me to grasp teachers' conception of the holistic and the analytic scales. In this regard, the questionnaire, which targeted the five teachers of oral expression, can be used as a first step to compare the holistic and the analytic scales. This questionnaire includes nine questions that aim at finding the way the teachers do assess their students' oral performance and understanding their conception of the holistic and the analytic scales.

D. The Comparative Design

The use of the audio recording and the questionnaires aims at making a comparison between the holistic and the analytic methods of assessment. Generally speaking, it is the teachers' scoring of the students' oral performances which serves as the first step towards making the comparison. Hence it is necessary to depict how this procedure takes place.

E. Scoring Procedure

After recording the twenty oral performances, I gave each tester a C.D that includes all the recorded speeches in order to assess them holistically and analytically. In fact, the students were classified into four groups (five students in a group). Each group was assessed by five testers according to one of the already specified tasks (interview, oral presentation, storytelling and picture description). Thus, I can reflect on the differences between the holistic and analytic scales by taking into account the variability of tasks to grasp whether the variability of the tasks shows any differences between the two methods of assessment or not.

1. Holistic versus Analytic Scoring

1.1 The Holistic Scoring

The teachers assessed the students' performances holistically. In fact, the table includes the number of the students, the different tasks according to which the students were tested , and the holistic scale that I took from Hughes' example of holistic scoring (1989). I chose this table, firstly, because it is clear in its classification of the levels ranging from "excellent" (18-20) to very poor (0-4), and secondly because it follows the same scoring procedures of scoring that are used in our Tunisian academic context, since all of the marks are between 0 and 20. Hence, the teachers would listen to the recorded performances and then give the marks according to six main criteria namely, excellent, very good, good, pass, weak or very poor.

1.2. The Analytic Scoring

The teachers scored the different performances while focusing on different aspects of speaking namely accuracy, fluency, pronunciation and content. The table of scoring does also include all the tasks and the students' numbers.

I took the example of scoring from the "Grading Scale of Oral Expression Interviews" which is followed in testing oral expression in the Higher Institute of Languages in Gabés. I selected this table because it efficiently reflects the ways by which the teachers of oral expression assess their students.

Hence, each level is given a (0.....5) scale which includes different criteria namely, awful (0), poor (1), o.k. (2), good (3), very good (4), and native like (5). In addition to that, each rubric of this analytic scale encapsulates the different parts of speaking. In fact, the grading scale illustrates the subcomponents of accuracy, fluency, pronunciation and content.

Firstly, accuracy is based on the extent to which a student's performance is grammatically consistent, which stands for the adequate use of subjects, verbs, tenses, prepositions etc... Secondly, fluency includes the use of lexis, idioms, synonyms, antonyms etc... This part of the analytic scale mainly requires the adequate use of vocabulary which enables the student to produce a coherent speech. Thirdly, the rubric of pronunciation is specified by the administrators of "the Grading Scale OE Interviews" as including: "past ed- endings, rhythm (not Fr. Syllable timed), correct stress on syllables and words, correct vowel pronunciation". Hence, it encompasses the bases of English pronunciation that should not be affected by Arabic or French pronunciation. Fourthly, the rubric of content is the one that is meant to assess the degree of coherence that the student's speech should adhere to. In fact, a coherent speech is based on some organised ideas that are related to a specific topic and on the extent to which a student is able to argue for or against an issue. Accordingly, the choice of this analytic scale comes as a result of its exhaustiveness, as it includes the basic subcomponents of speaking.

1.3 Reliability Measurement

a) Inter-rater Reliability

In this part, I made a comparison between the five raters' scoring, in order to determine the extent of agreement between them. As Cherry (2009) points out: "Another means of testing inter-rater reliability is to have raters determine which category each observation falls into and then calculate the percentage of agreement between the raters. So, if the raters agree 8 out of 10 times, the test has an 80% inter-rater reliability rate". I applied this method of comparison to calculating inter-reliability rates, because it was clear and concise. Furthermore, it was useful in reflecting on the similarities and differences between the five testers.

It has to be noted that the degree of agreement in the holistic scoring depends on the degree of agreement at the level of the six criteria that I already selected (Excellent, very good, good, pass, weak, very poor), so that, when two testers, for instance, give 12 and 14 to the same performance, I can state that both agree on the same criterion, since 12 and 14 come under the rubric of "good".

b) Intra-rater Reliability

In order to calculate intra-rater reliability rate, I used the same procedures that I selected for calculating inter-rater reliability, but with considerable modifications. In this regard, I directly compared the holistic and analytic scores given by the same rater, so that I was able to show the degree of agreement between the two sets of scores. Accordingly, I calculated the rate of similarity and difference of the two methods of scoring. Then, I made the comparison between them after analyzing the raters' outcomes.

F. The Questionnaire

Three testers out of five use the analytic scale to assess their students' oral performances, which reflects that most of the testers think that the analytic scale is more reliable than the holistic scale. Hence, four testers agree on the reliability of the analytic method of assessment.

Moreover, one tester uses both holistic and analytic scales when testing speaking, which shows his own conception of the two methods as distinct but complementary. In addition, the five testers agree on the subjectivity of the holistic assessment. This agreement confirms some of the researchers' depiction of the holistic scale as "impressionistic" (Cited in Mayo (1996).

Indeed, four testers believe that the analytic assessment has a positive backwash effect on teaching, which reflects that there is a tendency to assess the students' oral performances by evaluating each sub-component of speaking (pronunciation, vocabulary, message content, etc...) apart from the other components.

V. Analysis

After giving the 20 recorded oral performances to be evaluated by five testers holistically and analytically, I got different scores that were so useful in drawing a statistical framework. In this context, we can calculate **the Mean** in order to have a glance at the average given by each method of assessment.

A. The Mean of the Raw Scores

As table 1 shows, in the holistic assessment of the 20 oral performances, the testers' outcomes provide us with five means ranging from 8.5 (tester 1) to 10.3 (tester 4), i.e., the difference between the two means is of **1.8**. It is this difference which is important because we can deduce that despite the fact that the testers assessed their students using the same criteria, there is a difference at the average level between one tester and another tester.

TABLE 1:
RAW SCORES MEAN.

Testers	Holistic Scores Mean	Analytic Scores Mean
1	8.5	8.1
2	9.175	9
3	9.55	8
4	10.3	11.8
5	9.95	8.05
Total mean	9.495	8.99

As for the analytic scoring, the difference between the top mean (11.8, tester 4) and the down mean (8, tester 3) is of **3.8**, which reflects a considerable divergence between the two outcomes. Apart from that, the total means that are related to holistic and analytic scores are **9.495** and **8.99** respectively.

B. The Range of the Raw Scores

Concerning the holistic method of assessing speaking, the table of range (see Table 2) shows a difference between the testers' scores as far as the range is concerned. Indeed, the five ranges that are obtained are restricted between 12 and 8. Accordingly, the difference between the top range and the bottom range is **4**. This difference is significant in that it reflects the degree of agreement between the testers as they are assessing the same oral performances.

TABLE 2:
RAW SCORES RANGE.

Testers	Holistic Scores Range	Analytic Scores Range
1	11	9
2	10.5	10.5
3	8	9
4	8	7
5	12	9
Total range	9.9	8.9

Apart from that, the scores that are derived from the analytic scale yield a set of ranges that vary between 10.5 and 7. Hence, we obtain **3.5** as a range difference between the highest and the lowest scores. Moreover, the overall range that is specific to holistic scores is **9.9**, whereas the one that is related to analytic scores is **8.99**.

C. Reliability Criterion

In this section, we will take into consideration inter-rater reliability rate and intra-rater reliability rate, so that we can perceive the extent to which the two methods of assessment differ in terms of reliability as a criterion of making comparisons. As we have mentioned in the research design part, we will calculate the rate of agreement that is suitable for each type of reliability.

1. Inter-rater Reliability

The calculation of inter-rater reliability rate is based on the degree of compromise between the scorers as they give marks. Accordingly, the analysis of the agreement rate brings many useful results. All these results are shown in tables 3 and 4 (Appendices 1 and 2).

Table 3 illustrates the main results that were deduced after the calculation of the total agreement rate and inter-rater reliability rates according to the different tasks. The students were classified into groups according to task types.

2. Inter-rater Reliability as Deduced from the Holistic Scale

The calculation of the overall agreement rate in the holistic scale is of **62.22%.** In brief, we deduce that there is an acceptable extent of agreement between the raters, since it is over 50%, but we cannot ignore the 37.78% rate of disagreement, because the testers actually assessed the same set of oral performances.

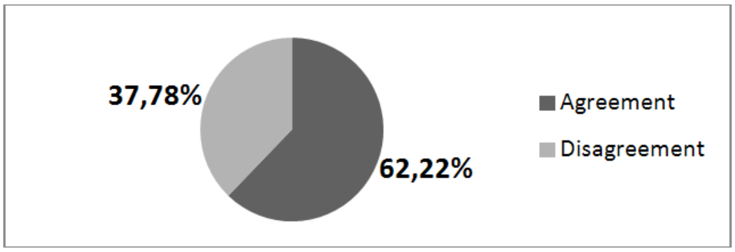

Figure 1: Inter-Rater Agreement in the Holistic Scale.

3. Inter-rater Reliability Rate According to Tasks

As far as interviews are concerned, we find that the total rate of agreement is 60%. In fact, we can consider this rate as low, since Stemler (2004) argues that: "A typical guideline found in the literature for evaluating the quality of inter-rater reliability based upon consensus estimates is that they should be 70% or greater"(p.3).

Apart from that, the assessment of the oral presentation reveals another set of reliability rates. It can be seen that the overwhelming percentage that characterizes the oral presentation rubric is 80%. Indeed, the calculation of the overall agreement rate reflects that there is a consensus over 73.33% of the scores. As for the picture description, the most noticeable rate of agreement is 80%. The five rates yield an overall rate of 62.85%. We have to mention that there is no full agreement between the five testers, as far as picture description is concerned.

Additionally, the reliability rates that are specific to storytelling task vary, with an overwhelming proportion of 60% that is specific to four performances. Hence, the overall agreement rate is 55%, which is lower than the overall rates that are specific to the other tasks.

4. Inter-rater Reliability According to the Analytic Scale

The overall accuracy agreement percentage is 48%, which is below the average of 50%. Accordingly, we can say that when the testers focused on the speech accuracy, they did not assess it in the same way, despite the fact that this criterion is actually based on clear rules (grammar and vocabulary rules).

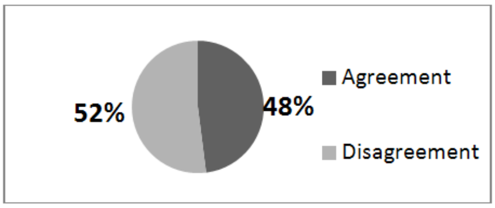

Figure 2: Inter-Rater Reliability According to Accuracy Criterion.

5. Inter-rater Reliability Rate According to Tasks

A look at table 5 reveals a difference of rates according to the four tasks. If we focus on the first task (the interview), we can remark that the majority of the raters agreed on 60% of the scores. Hence, the overall agreement rate is 56.66%, which is lower than Stemler's (2004) acceptable consensus estimates, which should reach at least 70%.

Furthermore, all of the rates that were given to the students who were assessed according to the oral presentation task are either 40% or 60%, yielding an overall percentage which is 46.66%.

TABLE 5:
INTER-RATER RELIABILITY AGREEMENT ACCORDING TO ANALYTIC ASSESSMENT: THE RUBRIC OF "ACCURACY".

Tasks	**Interview**	**Oral presentation**	**Picture Description**	**Storytelling**
Agreement rate	%56.66	46.66%	52%	45.75%
Total Agreement of Accuracy Scores 48%				

Next, the following task (picture description) reflects another set of agreement rates. It has to be noted that there is no agreement over the thirteenth student's speech accuracy, whereas the raters' feedback regarding the fifteenth performance reflects a higher agreement reaching 80%. After dealing with these two examples, it should be mentioned that there is a low reliability, since there is a huge difference between the two rates (0% and 80%). Accordingly, the overall agreement rate according to this task is 52%.

6. Inter-rater Reliability According to the Fluency Rubric

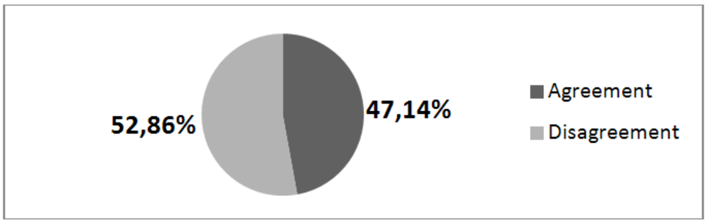

Figure 3:Inter-Rater Reliability According to Fluency Criterion.

In brief, the overall agreement over fluency scores is of 47.14%. This rate is low and demonstrates a lack of agreement between the raters as far as this skill is concerned. Hence, we can deduce that this low reliability can be traced back to the nature of fluency as a skill which is qualified differently by distinct raters.

TABLE 6:
INTER-RATER RELIABILITY AGREEMENT ACCORDING TO ANALYTIC ASSESSMENT: THE RUBRIC OF "FLUENCY".

Tasks	**Interview**	**Oral presentation**	**Picture Description**	**Story telling**
Agreement rate	51.42%	42.85%	50%	45%
Total agreement of Fluency Scores 47.14%				

The assessment of the first group of students according to the interview task resulted in different scores that affected their reliability. In this regard, if we look at table 6, I deduce that there are many examples showing that the same student was given extremely distinct scores. For instance, the marks that were specific to student 5 were as follows: 2, 1.5, 1, 3, 1. This difference is further reflected by the total agreement rate which is 51.42%. If we compare this rate to the one that is specific to the following task (oral presentation), we find that the second rate is even lower, reaching 42.85%.

Besides, all of the reliability rates that come under the rubric of the third task (picture description) are between 40% and 60%, resulting in 50% as a general agreement rate. In addition, the fourth task (storytelling) reflects a lower overall agreement (45%). If we look at the details, we find that the marks that are specific to student 16 are as follows: 2, 1.5, 3, 3, 1. Otherwise, there is a disagreement between the testers, especially between testers 3 and 4, on the one hand and tester 5, on the other hand.

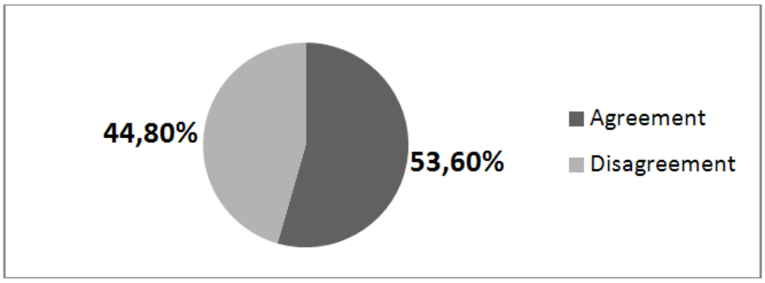

Figure 4: Inter-Rater Reliability According to Pronunciation Criterion.

After dealing with the percentage that is specific to each performance, we obtain 53.6% as a total agreement over pronunciation scores.

TABLE 7:
INTER-RATER RELIABILITY ACCORDING TO THE PRONUNCIATION RUBRIC

Tasks	Interview	Oral presentation	Picture Description	Story telling
Agreement rate	50%	54.28%	60%	50%
Total Agreement of Pronunciation Scores 53.6%				

If we concentrate on table 7, we can deduce that the testers agree on the same scores with varied rates. Apart from that, most of the rates that reflect raters' agreement range between 40% and 60%. However, we neither have any total disagreement (0%) related to the scores of any student, nor total agreement (100%).

Table 8 and figure 5 show that the total rate agreement of content scores is 51.66%. This percentage is acceptable as it is over 50%. If we look at the details that are given by the table, we can perceive that the agreement rates are mainly between 40% and 100%.

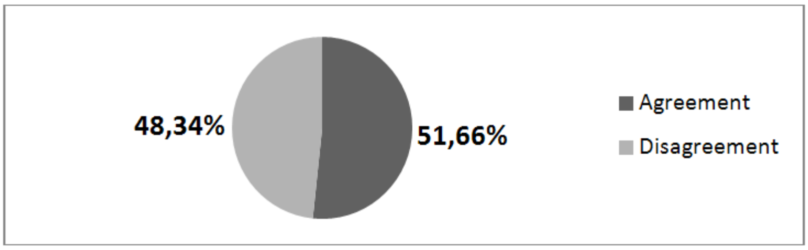

Figure 5: Inter-Rater Reliability According to Content Criterion.

Concerning the content rubric, the rates according to tasks range between 44% and 56.66%. The lowest rate (44%) is specific to the first task, where the raters agreed on 40% of the scores four times, which is considered as a low rate.

In the second task, the raters agreed one time on 80% of the scores. In the third task, the agreement rates are between 40% and 80%, resulting in 56% as an overall rate. In the fourth task, most of the rates are around 40%, yielding 50% as an overall rate.

TABLE 8:
INTER-RATER RELIABILITY ACCORDING TO CONTENT

Tasks	Interview	Oral presentation	Picture Description	Story telling
Agreement rate	44%	56.66%	56%	50%
Total Agreement of Content Scores 51.66%				

7. The Overall Agreement Rate of the Analytic Scale

As we have seen, the analytic scale encapsulates many criteria that can be analyzed in order to deduce the amount of reliability for each component. In this part, we will shed light on the total marks that emerged from this scale.

Accordingly, Table 6 (see Appendix 2) reflects the degree of reliability over the scores that are given to the twenty students.

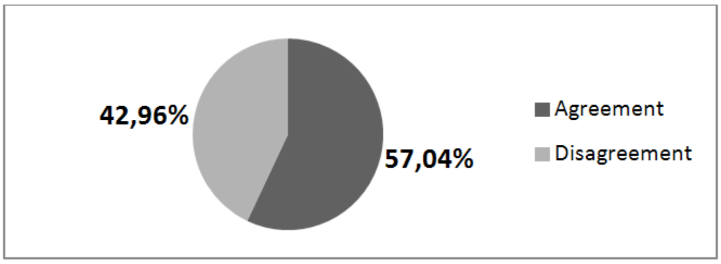

Figure 6: Analytic Scale Inter-Rater Reliability.

A scrutiny of table 6 reveals different rates that are specific to agreement according to tasks. In fact, the scores that are reported in the analytic scale are the composite scores of accuracy, fluency, pronunciation and content.

Both oral presentation and storytelling tasks engender the highest rate of agreement (60%). The task that reflects the least agreement rate (53.33%) is the interview. Moreover, the raters agreed on 55% of the scores that are under the rubric of picture description. In this regard, we have to mention that a full agreement (100%) occurs twice, since the testers gave almost similar scores to students 1 and 16. In addition to that, the testers agreed on 80% of the scores just five times, which shows the low total rate that characterizes the analytic scale, namely 57.037%. Thus, there is a variation according to tasks, but this variation does not hide the fact that the analytic scale does not guarantee an acceptable reliability rate (70%).

8. Intra-rater Reliability

If we focus on the first tester's outcomes (see table 9), we deduce that despite the fact that he gave the marks to the same performances, there was a considerable difference between the scores that were given according to the holistic scale and the ones that were given according to the analytic scale.

Apart from that, the total rate of agreement that is deduced from tester 2's outcomes is of **95%.** This rate is higher than tester 1's agreement rate. As a matter of fact, it seems that tester 2 does almost agree on the same set of scores holistically and analytically, which I consider as an exception because there is no other high rate of agreement.

After analyzing tester 3's holistic and analytic scores, we obtain **45%**, as a rate of agreement between the two scales. This rate is considered low, since it is below 70%.

TABLE 9:
INTRA-RATER RELIABILITY AGREEMENT.

Testers	Tester 1		Tester2		Tester3		Tester4		Tester5	
	H.S	A.S	H.S	A.S	H.S	A.S	H.S	A.S	H.S	A.S
Agr.	50%		95%		45%		65%		60%	

If we look at tester 4's feedback, we notice that the intra-rater agreement rate reaches 65%, which is above the average. However, we cannot ignore the 35% of disagreement which reflects a difference of rating.

9. Intra-rater Reliability Rate According to Tasks

A look at tables 10 shows that there is a difference between the rates according to tasks. In view of this, tester's 1 outcomes reflect distinct percentages. Hence, we find that tester 1 agreed on 40% of the scores that are both specific to interviews and storytelling tasks. It has to be noted, however, that the agreement rate reaches 60% as far as the oral presentation and picture description are concerned.

TABLE 10:
INTRA-RATER RELIABILITY RATE ACCORDING TO TASKS:

	Tester 1	Tester2	Tester 3	Tester 4	Tester 5	Total Agr.
Interview	40%	100%	20%	60%	60%	56%
Oral presentation	60%	80%	40%	60%	80%	64%
Picture description	60%	100%	80%	60%	40%	68%
Storytelling	40%	100%	20%	60%	60%	56%

Broadly speaking, the results imply a difference of intra-rater reliability rates according to tasks. As a matter of fact, the general agreement rate that is deduced from the scores of the first task (interviews) and the fourth task (storytelling) is 56%. The highest rate is related to the third task (picture description) which is 68%, almost reaching the acceptable rate of agreement (70%). Moreover, the rate of agreement that is specific to the second task (oral presentation) is 64%. In fact, we can say that intra-rater agreement rate depends on the task type and varies as the tasks vary.

VI. DISCUSSION

A. *The Mean*

The mean in the holistic method reflects that the holistic scale is more flexible than the analytic one. This view is further explained by Kozlowska et al (2005) who suggest that the holistic testing is based on a "criterion of intelligibility", whereas the analytic testing is based on error finding", since the testers focus on each part of the skill.

This goes in line with Smaoui (2009), who found that some aspects of holistic testing cannot be found in the analytic one. In fact, these results do not support some of the raters' viewpoints. The testers claim that the holistic scale is sometimes "unfair", since it is usually maintained by some researchers that the analytic scale is more objective than the holistic scale. Accordingly, the scorer can refer to the holistic scale and use it in assessing the performance together with the analytic scale. Thus, it seems that the analytic and the holistic scales can be used in a check and balance way.

B. The Range

We deduce that the calculation of ranges difference reflects a higher agreement between the raters in the analytic assessment than in the holistic one, but, we have to say that it is rather a slight difference because it is of **0.5** (the difference between 4 and 3.5). Moreover, the total range of the holistic scores is 9.9, whereas the total range of analytic scores is 8.9, which results in a difference of 1.

In fact, the holistic scale gives a deeper discrepancy between the students' levels. Accordingly, the holistic scale yields greater difference between the top and the bottom scores. In short, the holistic and the analytic methods of assessment are different in terms of range.

C. Holistic versus Analytic Scales

A look at the results that were deduced from the different scorers' rates according to the holistic scale shows that the overall agreement rate is 62.22%. In fact, this percentage reflects a low reliability since it is below the acceptable average of agreement (70%). It has to be noted that many researchers claim that any subjective assessment leads to low reliability (Hughes (1989)). Hence, it is clear that a holistic scale does not guarantee high reliability. Moreover, even at the level of tasks, we notice a considerable variation. For instance, the assessment of oral presentations yields an acceptable agreement rate (73.33%). This result can be explained by the fact that the raters of any oral presentation focus on a student who is speaking without any visual or written support. Thus, we can say that there are no clear cut parts in an oral presentation and we can notice the holistic aspect of this task. This stresses the fact that oral presentations are authentic tasks, when the students do not prepare them in advance.

In addition to that, the assessment of the other three tasks yields low reliability rates. It has to be noted that the more the students focus on parts of their speeches, the less reliability rate we get. In other words, we notice that when the students spoke freely in an oral presentation task, and without any interruption, the raters gave almost the same scores. This perception can be traced back to the fact that the testers did only focus on a specific speech which has a holistic aspect. If we compare oral presentation to the other tasks, we notice that the interviews, the picture description and storytelling tasks have a partitive aspect. Firstly, the picture description task necessitates that the students look at the different parts of the picture in order to depict it. Secondly, in the interview task, each student talked with an interlocutor (tester). Thus, the task can be divided into many pairs of speech. Thirdly, in the storytelling task, the students are required to focus on different words from which they can imagine a story. Hence, we deduce that there is a link between the partitive aspect of each task and the degree of reliability of the holistic scale.

After analyzing the data, we obtained different results specific to the analytic scale. As a matter of fact, we calculated reliability rates according to two distinct ways, firstly by determining the reliability rate of each criterion, and secondly, by finding the reliability rate by means of the total scores. Thus, we can say that the analytic scale is flexible since it enables us to interpret the data in two different procedures.

The study of the agreement rate of both methods results in many significant observations. Firstly, I infer that the study of analytic scale reliability requires more details than the holistic scale, because the analytic analysis of speaking is actually a scrutiny of all the components that make up this skill. Hence, I presume that the analytic scale offers a deeper observation related to calculating the degree of inter-rater reliability of accuracy, fluency pronunciation and content, which is not possible while analyzing the holistic scale data.

Secondly, I notice that inter-rater reliability in the analytic scale can be calculated in two different ways, either by dealing with the components of speaking or with the total (composite) scores that were given according to the scale, whereas, in the holistic scale, it is necessary to use the same set of scores in order to measure reliability. As a matter of fact, the analytic method seems to be more flexible than the holistic one, because it does not necessitate a single interpretation.

Thirdly, as far as inter-rater percentage is concerned, I find that its rate is higher in the holistic method (62.22%) than in the analytic one (57.037%). This perception reflects that the testers do agree more when using the rubrics that express their attitudes (excellent, good etc..) over the oral performance than when using the detailed analytic scale, as if there is a consensus between the teachers over the good and the weak speeches. Another interpretation of these results can be traced back to the nature of speaking itself, because when the listener in his/her daily life focuses on the oral performance, s/he cares more about the overall outcome than about its details. This observation does not support testers' conception that the holistic scale is less reliable than the analytic one, because their holistic assessment of the twenty oral performances revealed higher agreement than their analytic assessment.

Fourthly, at the level of tasks, I notice a greater variation of the reliability rate in the holistic scale than in the analytic one. Hence, I can state that the analytic scale is more suitable when dealing with different tasks, since it is useful when the students sit for oral exams that are based on different tasks. Thus, the variability of tasks deepened the differences between holistic and analytic approaches to assessment.

Fifthly, I confirm that the interpretation of the analytic results yields a deeper scrutiny of the different parts of speech, as well as of the way the testers focus on each part. It is clear that there is an unconscious weighting that characterizes each rater, and it is this concept which lessens scorers' reliability rates that are related to the analytic scale.

Eventually, all of these results reflect that the holistic and the analytic scoring of speaking are different in terms of inter-rater-reliability rate, which confirms our hypothesis that is based on the difference between the two methods of scoring.

It has to be noted that there is a common point between the two scales, which lies in yielding low reliability rates. We cannot ignore that the results reflect low agreement between the raters, because in both scales we did not attain the percentage of 70% of agreement except when we dealt with the holistic assessment of the oral presentation task (73.33% of agreement).

Thus, it is not frequent to obtain a high agreement between the raters as far as speaking is concerned. This fact can be explained by the concept of the unconscious weighting which we have dealt with in the previous parts of the research. Hence, the raters are different in their conception of what a good speech consists of.

VII. Conclusion

After dealing with the differences between the two methods, I notice that there are many differences between speaking assessment (as I dealt with it in this research) and writing assessment (as I read about it in the literature). Most of the researchers claim that the objective assessment (the analytic one) is more reliable when assessing writing. After dealing with speaking assessment in this research, I concluded that the holistic scale is more reliable than the analytic one.

Thus, we can see that there is a difference between holistic and analytic assessment of writing, on the one hand, and holistic and analytic assessment of speaking, on the other hand.

I think that making a comparison between these two productive skills in terms of holistic and analytic assessment is an interesting area of investigation, by which we can have a broader view of the use of these scales and provide a deeper analysis by setting a comparative design not only to show the differences between holistic and analytic scales, but also to highlight the discrepancies between speaking and writing.

Acknowledgements

I owe my deepest gratitude to my teacher Dr. Chokri Smaoui who inspired me to conduct this research.

Appendix 1

Table 5:
Inter-rater Reliability Agreement According to Holistic Assessment.

Students	Tasks	Tester 1	Tester 2	Tester 3	Tester 4	Tester 5	Total Agr.	Agr. According to tasks.
1	Interview	10	11	10	10	12	80%	60%
2	Interview	06	08.5	05	09	09	40% / 60%	
3	Interview	04	08	06	06	06	60%	
4	Interview	03	07	05	10	07	60%	
5	Interview	04	08	07	05	05	60%	
6	Oral presentation	13	13	13	13	15	100%	73.33%
7	Oral presentation	05	07	05	10	10	40% / 60%	
8	Oral presentation	08	07	10	10	11	80%	
9	Oral presentation	04	04	05	08	03	80%	
10	Oral presentation	06	08	09	10	08	80%	
11	Picture description	08	11	12	12	14	40% / 60%	62.85%
12	Picture description	10	10	09	11	05	80%	
13	Picture description	11	10	10	11	13	80%	
14	Picture description	14	10	10	13	12	40% / 60%	
15	Picture description	12	14.5	12	09	15	80%	
16	Storytelling	11	08	13	12	10	40% / 60%	55%
17	Storytelling	10	12	12	12	12	80%	
18	Storytelling	12	08.5	12	12	11	40% / 60%	
19	Storytelling	10	09	13	12	14	40% / 60%	
20	Storytelling	09	09	12	10	07	60%	
Total Agreement of Holistic Scores							62.22%	

APPENDIX 2

TABLE 6:
INTER-RATER RELIABILITY AGREEMENT ACCORDING TO ANALYTIC ASSESSMENT.

Students	*Tasks*	*Tester 1*	*Tester 2*	*Tester 3*	*Tester 4*	*Tester 5*	*Agreement*	*Agr. According to tasks*
1	*Interview*	*11*	*10.5*	*09*	*11*	*09*	*100%*	*53.33%*
2	*Interview*	*07*	*08.5*	*04*	*11*	*05*	*40% / 40%*	
3	*Interview*	*05*	*08*	*04*	*13*	*05*	*40%*	
4	*Interview*	*07*	*07*	*04*	*09*	*05*	*60%*	
5	*Interview*	*07*	*08*	*04*	*12*	*07*	*40%*	
6	*Oral presentation*	*13*	*12.5*	*13*	*16*	*11*	*80%*	*60%*
7	*Oral presentation*	*04*	*07*	*04*	*10*	*08*	*40% / 40%*	
8	*Oral presentation*	*08*	*05*	*09*	*09*	*09*	*80%*	
9	*Oral presentation*	*05*	*04*	*04*	*09*	*03*	*60%*	
10	*Oral presentation*	*05*	*07*	*07*	*13*	*08*	*60%*	
11	*Picture description*	*07*	*11*	*09*	*13*	*12*	*40% / 40%*	*55%*
12	*Picture description*	*07*	*10*	*08*	*11*	*06*	*60% / 40%*	
13	*Picture description*	*10*	*10*	*08*	*12*	*09*	*80%*	
14	*Picture description*	*13*	*10*	*08*	*13*	*11*	*60% / 40%*	
15	*Picture description*	*13*	*14.5*	*13*	*13*	*11*	*80%*	
16	*Storytelling*	*10*	*08*	*09*	*11*	*08*	*100%*	*60%*
17	*Storytelling*	*08*	*12*	*12*	*13*	*09*	*60% / 40%*	
18	*Storytelling*	*08*	*8.5*	*11*	*12*	*08*	*80%*	
19	*Storytelling*	*07*	*09*	*10*	*12*	*10*	*60%*	
20	*Storytelling*	*07*	*09.5*	*10*	*13*	*07*	*40% / 40%*	
Total Agreement of Analytic Scores							*57.037%*	

REFERENCES

[1] Alderson, J.C, C.Claphan & D.Wall. (1995). Language Test Construction and Evaluation. U.K: Cambridge University Press.

[2] Caldwell, B.M. (2009). From Babbling to Talking. Language and Learning. http://www.fisher-price.com/en_US/parenting-articles/language-and-learning/from-babbling-to-talking (accessed 4/2/2016).

[3] Cherry, K. (2009). What Is Reliability? http://psychology.about.com/od/researchmethods/f/reliabilitydef.htm (accessed 2/6/2015).

[4] Fasold, R.W. (1990). The Sociolinguistics of Language (vol 2). U.S.A: Blackwell Publishers.

[5] Garman, M. (1990). Psycholinguistics. Cambridge Textbooks in Linguistics. U.K: Cambridge University Press.

[6] Harris, D.P. (1969). Testing English as a Second Language. U.S.A: Georgetown University.

[7] Hornby, A.S. (2000). Oxford Advanced Learner's Dictionary of Current English. U.K: Oxford University Press.

[8] Hughes, A. (1989). Testing For Language Teachers. U.K: Cambridge University Press.

[9] Kozlowska, J.S, J. Frankiewicz, M. Nowacka & L. Stadnicka. (2005). Assessing Assessment Methods: On the Reliability of Pronunciation Tests in EFL. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.546.5252&rep=rep1&type =pdf (accessed 5/1/2016).

[10] Kumar, R. (2005).Research Methodology: A Step-by-Step Guide for Beginners. Australia: Sage Publications.

[11] Levelt, W. (1989). Speaking: From Intention to Articulation. England: Massachusetts Institute of Technology.

[12] Luoma, S. (2004). Assessing Speaking. U.K: Cambridge University Press.

[13] Mayo, P.G. (1996). The Reliability of the Holistic Method When Grading Language Essays. Del Pais Vasco University. *Cuadernos de Filologia Inglesa*. 51-62.

[14] Mertler, C. A. (2001). Designing Scoring Rubrics for your Classroom. *Practical Assessment, Research & Evaluation*, 7(25), 1-8.

[15] Nakamura, Y. (2004). A Comparison of Holistic and Analytic Scoring Methods in the Assessment of Writing. The Interface between Interlanguage, Pragmatics and *Assessment: Proceedings of the* 3^{rd} *Annual JALT PAN SIG Conference.* http://jalt.org/pansig/2004/HTML/Nakamura.htm (accessed 2/2/2015).

[16] Richards. J.C. (2008). Developing Classroom Speaking Activities: From Theory to Practice. file:///C:/Users/pctoshiba/Downloads/developing-classroom-speaking-activities.pdf (accessed 6/7/2016).

[17] Scholes, R.E. (2006). The Nature of Narrative. New York: Oxford University Press.

[18] Smaoui,C. (2009). The Silence/Talk Interplay: Fluency in Tunisian Learners' Oral Production. In Guirat, M and Triki, M. (eds.), *Silence.* Tunisia: University of Sfax. Faculty of Letters and Humanities, Grad.

[19] Stemler, S.E. (2004). A Comparison of Consensus, Consistency and Measurement Approaches to Estimating Interrater Reliability. Yale University. *Practical Assessment, Research and Evaluation Journal.* V9, N4. 1-11.

[20] Xi, X & P.Mollaun. (2006). Investigating the Utility of Analytic Scoring for the TOEFL Academic Speaking Test (TAST). TOEFL iBT Research Report. http://s3.amazonaws.com/academia.edu.documents/30193670/rr-06-07.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1488955167&Signature=j2SSmKbSxuiIK%2B9aZKX%2BcdJ%2BSf8%3D&response-content-disposition=inline%3B%20filename%3DInvestigating_the_utility_of_analytic_sc.pdf (accessed in 15/01/2017).

Maha Ounis was born in Gabés, Tunisia. She holds an MA in Theoretical and Applied Linguistics from the Faculty of Letters and Humanities of Sfax (FLSHS), Tunisia. Her research interests include Language Testing, Teaching Methods and Second Language Acquisition.