The Impact of Analytical Grading Criteria on Holistic Grade in the Speaking Assessment

By Ulker Ibrahimova,
ADA University,
Azerbaijan

Abstract
With the advance of communicative approach in language teaching and learning, the same tendencies started to be observed in another inseparable area of education: assessment. The move has been quite difficult especially in the areas of the assessment of productive language skills, writing and speaking, owing to the openness to subjective interpretation. Notwithstanding, the move to more interactive formats in their assessment has been steady. The most popular for the today is the format of pair or group interviews in the assessment of oral language performance. This move has been naturally accompanied by the growing demand for more research as to the effect of the format on the validity and reliability of the test and its results. The current paper aims to contribute to the research findings by exploring the relationship between the various analytical grading criteria and the holistic grade as well studying the impact of the criterion of interaction on the global holistic rate.

Keywords: paired and group speaking assessment, pair interviews, interaction, analytical grading criteria, correlation, impact on the holistic rate
Introduction

In the area of foreign language assessment, more and more attention is paid to the assessment of productive skills – writing and speaking - in a way that would resemble target language use situation. In assessment of speaking, this meant shift from the format predominantly focusing on dictation and pronunciation in the early 20th century to the assessment of ability to interact and perform in the times of Second World War (Fulcher, 2003). This was taken further by Canale and Swain’s (1980) study of critical importance that emphasized the necessity to consider “communicative competence” (p.1). This line of argumentation opened up venue for questioning whether the competence lies within the realm of the possibilities of an individual language user, or whether it is a compound construct, or rather, co-construct. The same doubt was expressed by McNamara (1997) who believed the models employed at the time both in teaching and assessment devalue the interaction component and the individual’s performance in interaction process.

Therefore, in the sphere of language learning and teaching, with a growing demand towards more authentic assessment (Taylor, 2003), many educational institutions have opted for inclusion of paired and group interviews in the assessment of the speaking skill (East, 2015; Galaczi, 2014). The tendency is understandable since the pair/group format replicates pair/group-work classroom activities, is more practical in terms of expenditures of human resources and time, and more authentic as the tasks simulate certain target language use situations that the test-takers are likely to encounter outside the educational setting (Ducasse & Brown 2009). However, there are also controversies associated with the format. First of all, they concern the impact the partner or the peer’s age, gender, the degree of familiarity, and level of language proficiency may have on a test-taker’s speaking performance. Indeed, there have been an extensive number of research into peer-to-peer interaction that have cautioned against the effect a peer may have on the test-taker (Foot, 1999; O’Sullivan, 2002; Fulcher, 2003). They were concerned primarily with the possibility that the partner’s personal characteristics may affect the test validity. However, more recent research findings seem to suggest that the advantages exceed the possible limitations of the paired/dyadic format (Norton, 2005). One of the biggest advantages the proponents of the pair format name is the possibility of test-takers’ embarrassment and even intimidation by the examiner’s language proficiency or even the authority in a traditional examiner-test-taker format (Brooks 2009; Birjandi 2011), that is, when the “partner” for the test-taker is the examiner himself/herself.

Indeed, the extensive body of research targets more the process of the pair/group format of the speaking assessment, and the effect of peers’ characteristics on each other and the speech samples produced. Yet, the research studying the assessment itself and the interplay of speaking assessment criteria is limited. This is confirmed in the review of empirical studies targeting the second/foreign language oral testing conducted by Sandlund, Sundqvist and Nyroos (2016) who...
have also found that the bulk of the research concerns oral proficiency or placement tests, and much less so, the other test types such as diagnostic, or achievement (Brown & Abeywickrama, 2010). Therefore, the current paper aims to contribute to the research of speaking assessment of achievement test and the study of its reliability in terms of the impact the analytical grading criteria may have on one another, and specifically, on the overall/holistic rate.

The following questions guided the study:

1) How are the grades distributed in relation to each of the criteria?
2) How does each of the criteria of the analytical scale used by the assessor impact the holistic rate given by the examiner?
3) Is there any significant impact of the criterion of interaction on the holistic rate of the examiners?

Literature Review

Following the concept of communicative competence (Canale and Swain, 1980), the concept of interactional competence was taken up by Kramsch (1986, p. 367) in the sense that has dominated the area of language teaching, learning and assessment:

Whether it is a face-to-face interaction between two or several speakers, or the interaction between a reader and a written text, successful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or “sphere or inter-subjectivity” that is built through the collaborative efforts of the interactional partners.

The emergence of these concepts was marked by the gradual shift from the focus on an individual and his/her speech production to the collaborative “co-construction” (May, 2010) that entails “interactional processes, including collaboration, coordination, and cooperation” (Jacoby and Ochs, 1995, p. 171). Such processes were later called interactive practices (Hall, 1995) or discursive practices (Young, 2011). The word “practices” referred to its repeated occurrence in specific culturally bound contexts, yet not limited by its time or place. It also meant that interactional competence is not a competence that exists per se, by itself, is not “the knowledge or the possession of an individual person, but is co-constructed by all participants in a discursive practice” (Young, 2011, p. 428). This understanding of interactional competence as an integral part of a person’s linguistic competence, which is at the same time circumscribed by the context shared and co-created with an interlocutor called for a new format of teaching language skills, and naturally, for a new format of speaking assessment. The change was relatively easy to implement in teaching and learning through interactive tasks and pair-work, group-work as new dominant ways of classroom management. Yet, in the area of language assessment, the attempt to tap into...
interactional competence was long suspended by high-stakes oral proficiency interviews (OPIs) (Sandlund, Sundqvist, Nyroos, 2016). Even though OPIs are extensively employed in Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS), they have been extensively criticized. First, they are considered unnatural and not resembling real-life conversations, second, the test-takers’ performance may well be inhibited by the examiner’s personality, authority, or language level (Brown, 2003).

In an attempt to create opportunities for extended samples of individual speech and co-constructed oral performance, pair or group interview format has emerged. It is considered as more beneficial as it is more authentic, relatively replicating classroom interaction and real-life situations, and can be more beneficial since the test-takers are paired with the peers of relatively similar characteristics (Son, 2016). Moreover, frequent criticism that the interlocutor’s language proficiency may inhibit the test-taker’s performance was found ungrounded in Davis’ study (2009). Yet, it is considered quite problematic in terms of assessment as the test-takers jointly participate in the construction of interaction, yet are assessed individually (Sandlund & Sundqvist 2011). Still, a number of research have found that in terms of the perception of the involved parties, both language educators and test-takers generally find the pair/group speaking test format to be more effective, authentic and engaging (East, 2015).

Methodology

Participants

The participants of the study are students of foundation level of intensive English for Academic Purposes Program (EAPP) at ADA University. They are all within the same age range, 16-18, of different gender, and various majors. At the onset of the academic year, the newly enrolled students take placement tests based on the results of which they are consequently placed in one of the following levels: Elementary, Pre-Intermediate, Intermediate and Upper-Intermediate. Students study for 8 weeks to complete each level. At the end of every eight weeks, the students take a speaking achievement test that is supposed to measure their oral language proficiency and interactive competence.

For the purposes of this study, the results of 98 Elementary, 101 Pre-intermediate, 100 Intermediate, and 84 Upper-intermediate level students are analyzed. The results of different years of speaking assessment are randomly taken, from 2015 to December 2018.

The participants that received the same score for each of the analytical scoring criteria and the overall holistic rate were not considered as their results would make no significance to the study of the posed research questions.
Description of the Speaking Exam

Test format

The speaking test is administered as an end-of-course achievement exam at four levels: elementary, pre-intermediate, intermediate, and upper-intermediate. The format of the exam is uniform across the levels; however, the allotted time and the difficulty and content of the tasks varies across the levels. The speaking achievement test consists of the following parts:

- a short introduction by the examiner which includes greetings, introduction of self and the assessor, the short description of the test and timing (ungraded part).

- Task 1: individual interview with three short open-ended questions asked at a stretch to each test-taker.

  Since this is an achievement test, the questions are based on the topic range covered within a course. The examiner is provided with a menu of questions so that each test-taker receives a different set of questions. The questions target different topics, skills, and grammar points to elicit representative speech sample. Regarding the timing, two minutes are allotted per test-taker at elementary level, and three minutes - from pre-intermediate level and above.

- Task 2: pair or group interaction
  1 minute: thinking time during which test takers can take notes but cannot talk to each other
  3 minutes (pair): interaction
  4 minutes (group): interaction
  For this part, the test-takers choose a card from the two stacks of visual and situational prompts laid in front of them: either a set of pictures or a set of written prompts. The tasks are based on the course objectives while the topics are based on the content of the respective levels.

Test Administration Procedure

The test is conducted by two EAPP native and non-native English instructors: an examiner that guides the exam and an assessor that does not interfere in the course of the exam and only grades the test-takers. To ensure stronger interrater reliability, the examiner and the assessor are not the teachers who have taught the group during the preceding eight-week period in which case they would know the test-takers and their overall language proficiency and may be affected by this knowledge. For better intra-rater reliability, the assessor and the examiner swap their roles after
every 5 or 6 pairs of test-takers (The number may vary depending on the overall number of pairs). The students are from the same group randomly placed in pairs.

The examiner sits facing the test-takers and instructs and guides the speaking exam. It is also the examiner’s responsibility to keep the time. In case the test-takers do not understand a question during task 1, the examiner may repeat the question or change the question. This does not affect the test-taker’s grade in any way. The rationale behind the absence of judgement in this case is due to the nature of comprehension break-down in real life. While in real-life an interlocutor has a chance to ask for clarification or reformulate, it is impossible to allow for this in a test situation without compromising test reliability. Hence, the examiner is allowed to change the question but not reformulate or explain for clarity. In task 2, in case of dominance of a test-taker, the examiner can encourage the more passive interlocutor with a wave of a hand, but not verbally, again to ensure the uniformity of the test administration across the board. Upon the completion of each pair or group speaking test, the examiner is required to assess the test-takers based on the holistic assessment scale, named Global Achievement Scale (Henceforth: GAS).

The assessor only greets the test-takers when the examiner introduces them. The assessor silently sits diagonally behind the test-takers, invisible to them so that the sight of an assessor taking notes on their performance does not intimidate the test-takers. The assessor’s task is to only assess the test-takers’ produced speech based on the analytical grading scale.

The instructors are regularly trained to conduct the speaking test and have moderation sessions once in every eight weeks for calibrating the assessment process. Video recordings of previous speaking tests are used for moderation sessions.

**Instruments**

Four sets of speaking tests were used for the speaking assessment: one for each level (Elementary, Pre-intermediate, Intermediate, and Upper-intermediate). Two types of assessment scales were used to collect the data: Analytical assessment scale used by assessors (see App 1) and Global Achievement scale – a holistic assessment rubric used by examiners (see App 2).

Analytical scale is a 20-point rubric measuring 5 criteria of discourse management, grammatical range and accuracy, vocabulary range and accuracy, intonation and pronunciation, and interaction.

GAS is a 4-point holistic scale, the aim of which is to assess “the candidate’s overall effectiveness in dealing with the tasks in the two separate parts of the Speaking test. The global mark is an independent, impression mark that reflects the assessment of the candidate’s performance from the examiner’s perspective” (App 2).
The overall rate referred to as “Total” (App 1) is calculated based on the sum of the assessor’s analytical scale grade and examiner’s holistic scale grade.

Data Analysis

The statistical analysis of the scoring provided both by the examiners and the assessors was employed. The first research question concerned the distribution of grades across the grading criteria; for this purpose, the descriptive analysis is used, and the bar charts are produced for each of the criteria. To address the second research question, that asked how each of the analytical criteria impacts the holistic global grade given by the examiner, linear regression analysis was used to reveal the relationship between the analytical criteria and the examiner’s global grade.

For exploring the answer to the third research question if there is any significant impact of interaction on the holistic grade, bivariate Pearson correlation coefficients of the analytical criterion of interaction and the global holistic grade were conducted to examine the relationship between the two scores.

Results

Distribution of Grades across Assessment Criteria

The descriptive statistical analysis revealed the steady pattern in distribution of grades alongside the four-point gradation in each of the criteria. Each of the criteria has very few numbers of “very poor” scores: discourse management (4.1%), grammatical range and accuracy (6 %), and vocabulary range and accuracy (3.4%), and interaction (2.3%). The criteria of intonation and pronunciation is the least amongst the analytical criteria that was graded with the lowest scores with 1.3%. This can be accounted for by the descriptor of the related level in the analytical scoring scale (App 1) which specifies the very poor performance as “mostly unintelligible”.

Table 1

Descriptive Analysis of the Distribution of Grades for Analytical Grading Criteria
### Analytical criterion of discourse management

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very poor</td>
<td>16</td>
<td>4.1</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>struggling</td>
<td>76</td>
<td>19.7</td>
<td>19.7</td>
<td>23.8</td>
</tr>
<tr>
<td>good</td>
<td>158</td>
<td>40.9</td>
<td>40.9</td>
<td>64.8</td>
</tr>
<tr>
<td>excellent</td>
<td>136</td>
<td>35.2</td>
<td>35.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>386</td>
<td>100.0</td>
<td></td>
<td>100.0</td>
</tr>
</tbody>
</table>

### Analytical criterion of grammatical range and accuracy

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very poor</td>
<td>23</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>struggling</td>
<td>111</td>
<td>28.8</td>
<td>28.8</td>
<td>34.7</td>
</tr>
<tr>
<td>good</td>
<td>208</td>
<td>53.9</td>
<td>53.9</td>
<td>88.6</td>
</tr>
<tr>
<td>excellent</td>
<td>44</td>
<td>11.4</td>
<td>11.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>386</td>
<td>100.0</td>
<td></td>
<td>100.0</td>
</tr>
</tbody>
</table>

### Analytical criterion of vocabulary range and accuracy

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very poor</td>
<td>13</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td>struggling</td>
<td>94</td>
<td>24.4</td>
<td>24.4</td>
<td>27.7</td>
</tr>
<tr>
<td>good</td>
<td>158</td>
<td>40.9</td>
<td>40.9</td>
<td>68.7</td>
</tr>
<tr>
<td>excellent</td>
<td>121</td>
<td>31.3</td>
<td>31.3</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>386</td>
<td>100.0</td>
<td></td>
<td>100.0</td>
</tr>
</tbody>
</table>
Analytical criterion of intonation and pronunciation

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very poor</td>
<td>5</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>Struggling</td>
<td>32</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>Good</td>
<td>192</td>
<td>49.7</td>
<td>59.3</td>
</tr>
<tr>
<td>Excellent</td>
<td>157</td>
<td>40.7</td>
<td>99.3</td>
</tr>
<tr>
<td>Total</td>
<td>386</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

analytical criterion of interaction

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very poor</td>
<td>9</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Struggling</td>
<td>58</td>
<td>15.0</td>
<td>17.4</td>
</tr>
<tr>
<td>Good</td>
<td>151</td>
<td>39.1</td>
<td>56.5</td>
</tr>
<tr>
<td>Excellent</td>
<td>168</td>
<td>43.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>386</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The descriptive analysis also revealed the highest concentration of grades across most analytical criteria in the third level of descriptors specified as “good”, with highest concentration of 53.9% for grammatical range and accuracy (Table 1). The figures for the criterion of interaction differ in the highest concentration of grades marked as “excellent”, with 43.5%.

With regards to the distribution of grades for holistic assessment (Table 2), the highest concentration of grades coincides with the overall pattern observed for analytical criteria, that is, for the level of “good”, with 40.7%.

Table 2

*Distribution of Grades for Global Holistic Grade*

*Global holistic grade*
The lowest concentration as expected is observed for “very poor” performance, with 2.6%. These results can be justified by the very nature of the test type. Achievement tests by virtue of being course- and objective- oriented normally produce many more results of higher than average, and very few low scores.

**Relationship between Analytical Criteria and Global Holistic Grade**

The results of the regression analysis revealed that there is statistically significant relationship between three of the analytical criteria and the global holistic scale. The analytical criterion of discourse management ($p < 0.000$), of vocabulary range and accuracy ($p < 0.008$), and of intonation and pronunciation ($p < 0.009$) seem to have a significant effect on the global holistic grade.

<table>
<thead>
<tr>
<th>Mode 1</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant) analytical criterion of discourse management</td>
<td>.567</td>
<td>.187</td>
<td>.342</td>
<td>.050</td>
</tr>
<tr>
<td>(Constant) analytical criterion of discourse management</td>
<td>6.855</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
On the other hand, the p-values for analytical criteria of grammatical range and accuracy (p<0.062) and that of interaction (p<0.073) seem to suggest that there is no significant relationship between these two criteria and the global holistic grade.

The Impact of the Criterion of Interaction on the Global Holistic Grade

The results of the Pearson Correlation coefficients indicate there is no significant correlation between the criterion of interaction and global holistic scale. The correlation coefficient of 0.000 reveals no relationship between the two variables even though it was believed to be essential, if not decisive in the formation of the examiner’s holistic assessment. The assumption was based on the notion of “intersubjectivity” introduced by Kramsch as the ability to construct interaction “within the shared internal context” (1986). The supposition was based on the assumption that the examiners given the opportunity to assess the test-takers globally, with an impression mark based on a holistic scale, would subconsciously regard the criterion of interaction, or co-construction of interchange, as more important. However, the statistical analysis seems to reject the assumption.

Table 4

| analytical criterion of grammatical range and accuracy | .108 | .058 | .099 | 1.871 | .062 |
| analytical criterion of vocabulary range and accuracy | .137 | .051 | .140 | 2.675 | .008 |
| analytical criterion of intonation and pronunciation | .152 | .058 | .126 | 2.616 | .009 |
| analytical criterion of interaction | .088 | .049 | .086 | 1.800 | .073 |

a Dependent Variable: Global holistic grade
Correlations

<table>
<thead>
<tr>
<th>analytical criterion of interaction</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
<th>Global holistic grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global grade</td>
<td>Pearson Correlation</td>
<td>Sig. (2-tailed)</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.389(**)</td>
<td>.000</td>
<td>386</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>.000</td>
<td>386</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>386</td>
<td>386</td>
<td></td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).

This finding proves the null hypothesis that there is actually no significant impact of the analytical criterion of interaction upon the global holistic grade. In fact, the regression analysis indicated a strong relationship between three other criteria, in particular, the analytical criterion of discourse management and global holistic grade.

Conclusion

The distribution of grades across each of the analytical criteria and the holistic grade was studied, revealing the lowest scores in “very poor” (marked 1 in the rubric; see App 1) and the highest concentration in “good” (marked 3 in the rubric; see App1) levels of descriptors. The criterion of interaction stands out with highest concentration in “excellent” (marked 4) score.

Furthermore, the relationship between the analytical criteria and the global holistic grade was explored. The analysis indicated that there is strong relationship between the criteria of discourse management, vocabulary range and accuracy, intonation and pronunciation, on the one hand, and the holistic grade, on the other. Surprisingly, the analysis of the remaining two analytical criteria of grammatical range and accuracy and interaction did not reveal any statistically significant relationship and effect on the holistic grade.
The possibility of the impact of the analytical criterion of interaction upon the holistic grade was rejected by the results of the correlations of the figures for both variables. The null hypothesis of no significant relationship between the two criteria was proven.

However, there are limitations to the study to be acknowledged. First, the analysis of the scoring results per each of the Elementary, Pre-intermediate, Intermediate, and Upper-intermediate levels would be more desirable with more representative sample numbers. The higher representative sample will yield stronger evidence as to the interplay of the criteria and the global holistic grade. Secondly, a post-test interview of questionnaire with the examiners who are responsible for the holistic assessment could shed light to the inner judgement and rationale that the examiners employ for their impressionist scoring of the students’ overall speaking performance. Furthermore, for future studies, the analysis of the stand-out cases when the holistic grade differed from all the other analytical scores could be conducted to reveal the cases that do not follow the general pattern. It would be worthwhile to follow such cases in particular with interviews of the examiners to explore the mechanisms and justifications of such different assessment.

Reference List

Birjandi, P. (2011). From face-to-face to paired oral proficiency interviews: the nut is yet to be cracked. *English Language Teaching* 4. 169–75


Fulcher, G., Davidson, F., Kemp J. (April 2010). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*


Appendix 1

Analytical Assessment Scale

<table>
<thead>
<tr>
<th>Criteria</th>
<th>G</th>
<th>Descriptors</th>
<th>Student Name/Surname:</th>
<th>Ave</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse management</td>
<td>4</td>
<td>Produces responses extended well beyond short phrases, with frequent use of long sentences. Contributions are always relevant, with only occasional repetition. Uses a good range of cohesive devices. The task is fulfilled well.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Produces responses with a blend of short phrases and some long sentences. Although there is some repetition, this does not impact the flow of ideas. Uses a range of cohesive devices. The task is mainly fulfilled.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Produces some extended phrases and with occasional hesitation. Repeats information and digresses occasionally. Uses a limited range of cohesive devices. Organization is occasionally unclear. Parts of the tasks are not fulfilled.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Responds with frequent pauses. Often repeats information and has very limited use of cohesive devices. The tasks are not fulfilled to a satisfactory standard. Some L1 is used.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammatical range and Accuracy</td>
<td>4</td>
<td>Shows a very good degree of control of grammatical forms with very few errors made.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Shows a relatively good degree of control of grammatical forms with occasional errors.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Shows some control of grammatical forms, attempts the occasional complex grammatical form, although it may be wrong, and errors are frequent.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Shows very limited control of grammatical forms with no attempt at complex grammatical forms; mostly uses simple forms.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary range and accuracy</td>
<td>4</td>
<td>Uses a wide range of usually appropriate vocabulary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Uses a sufficient range of appropriate vocabulary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Uses only occasionally new vocabulary with a very limited range, frequently repeating.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Uses very limited vocabulary and phrases and often inappropriately</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intonation and pronunciation</td>
<td>4</td>
<td>Has good control of pronunciation and intonation at utterance level with a natural flow at sentence level. Has good awareness of all phonemes.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Has generally good control of pronunciation and intonation at word and sometimes a natural flow at utterance level. Has difficulty with some phonemes.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Limited control of pronunciation. Intonation can often reduce full understanding of meaning leading to some unintelligibility. Frequent mispronunciation of words and frequent mispronouncing of phonemes.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Has very limited control of pronunciation and intonation. Phoneme use and intonation prevents full understanding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction/Turn-taking</td>
<td>4</td>
<td>Effectively initiates and maintains interaction and develops exchanges using learnt strategies. Requires no prompting and support and maintains good balance of speaking time.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Initiates interaction reasonably effectively with some natural exchanges using learnt strategies. Occasional prompting required or at times tends to dominate.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Initiates interaction infrequently and manages to maintain only very simple exchanges. Requires prompting and support or is too dominant</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Has difficulty maintaining simple exchanges. Requires a lot of additional prompting and support.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Examiner’s global rate (4/3/2/1)

TOTAL (out of 24 points)
Appendix 2

GLOBAL ACHIEVEMENT SCALE: Intermediate

Aim: This scale refers to the candidate’s overall effectiveness in dealing with the tasks in the two separate parts of the Speaking test. The global mark is an independent, impression mark that reflects the assessment of the candidate’s performance from the examiner’s perspective.

4: Effective command of spoken language in familiar situations. Can organize extended discourse with only occasional lack of coherence, inaccuracies and usage. Maintains a flow of language, with occasional hesitation while searching. Pronunciation is easily understood. Maintains and develops interaction, links contributions and negotiates an outcome.

3: Usually effective command of spoken language, with some hesitation and repetition while searching. May sometimes use inaccurate and/or inappropriate linguistic resources to express ideas; produces extended discourse that may sometimes be illogical. Pronunciation sometimes puts stress on listener. Maintains and develops interaction and negotiates and outcome.

2: Ineffective command of spoken language in familiar situations. Struggles to organize extended discourse and produces responses that often lack coherence, with frequent inaccuracies and inappropriate usage. Keeps the interaction going, but does not develop ideas. Needs support to interact. Pronunciation is difficult to understand and heavily influenced by L1.

1: Unable to produce extended discourse or coherent sentences. Cannot keep the interaction going.