# The Effects of Visual Input on Scoring a Speaking Achievement Test

**Jorge Beltrán**[1]
*Teachers College, Columbia University*

## ABSTRACT

In the assessment of aural skills of second language learners, the study of the inclusion of visual stimuli has almost exclusively been conducted in the context of listening assessment. While the inclusion of contextual information in test input has been advocated for by numerous researchers (Ockey, 2010), little has been said regarding the *scoring* of speaking tests, which also involves raters' listening comprehension. This study is designed to identify the possible variation in the scoring of speaking test performance when the speech samples to be scored are presented in either audio-only or audio-visual format. A group of raters were first asked to score a set of audio-only speech samples from an achievement speaking test consisting of one monologic task through an online platform. Weeks later, they scored the same samples presented in audio-visual format. Scores from both scoring sessions were compared. Findings suggest that the inclusion of visual stimuli may not result in significant effects on assigned scores or internal consistency. Yet, given the raters' reported preference of using the audio-visual format, the study results call for further exploration of the positive effects of delivery methods on rater effect.

## INTRODUCTION

Speaking assessment has been regarded as one of the most challenging types of assessment in second and foreign language teaching and learning. It involves the development of an elicitation instrument, a rating scale, and the training and norming of interviewers and/or raters. In addition, whether live or recorded, speaking assessment involves raters having to rely on their listening skills and often times their short-term working memory, as opposed to raters of written samples, who always have a document to rely on (Ginther, 2013).

Much of the complexity of speaking assessment comes from the standpoint of construct definition, given the multiple views on what speaking entails and how this is to be assessed (Fulcher, 2003). The tasks of a speaking test should be designed based on an ability model, that is, a speaking ability construct, so that the selected tasks elicit the type of language that is representative of the target language use domain, and which is to be assessed based on a rating scale that is aligned with the expected linguistic behavior (Luoma, 2004).

[1] Jorge Beltrán is a student of the Ed. D. in Applied Linguistics at Teachers College Columbia University in the language assessment track. His current research interests include learning-oriented assessment, assessment of speaking, construct validation, and performance-based assessment. He can be reached at: jlb2262@tc.columbia.edu

Another important aspect to consider in the assessment of speaking ability is the intended population of test takers, so that adequate criteria for rubric development and test task design are set (O'Sullivan, 2014). A number of variables may affect test taker performance. Examples of such variables may be the "interlocutor, task format, task topic, the examiner, and the previous knowledge of the scoring system" (O'Sullivan, 2014, p. 159). Moreover, rater(s) may also be influenced by test taker characteristics, the task itself, or the scoring system (O'Sullivan, 2014).

In order to avoid threats to validity, method variance should be kept to a minimum. One of the main causes of method variance is rater effects, which add construct-irrelevant systematic variance depending on the particular tendencies of a rater (Eckes, 2005). Rater training and norming are, therefore, fundamental in the validation of speaking assessments since they allow for the development of effective scoring standards through the deeper understanding of test rubrics. Nonetheless, bias can permeate even after norming sessions, and undesired variability may occur. Then, what could be some of the factors that lead to method variance in relation to rater effect? Would, for instance, presenting raters with the video recording of a test performance instead of audio recording only have an effect on raters' scoring performance? These questions seem particularly important in a time when remote rating though computer mediated systems increases in popularity. In light of this, the current study sought to determine whether the mode of delivery of speech samples (i.e., audio-only or audio-visual) has a systematic effect on scoring, and whether raters would report a preference for either one. In the following section, a brief review of variables influencing rater performance will be revisited.

## LITERATURE REVIEW

By definition, tests of speaking ability are considered subjective in nature since they require the judgment of a human rater (Carr, 2011). In spite of the numerous efforts that are made to enhance objectivity in speaking assessments, intra- and inter-rater consistency are aspects of the scoring process that need to be monitored given their fundamental role in score assignment (Czepes, 2009). Thus, one of the most important steps in enhancing reliability of a speaking test is the standardization of its scoring procedures and alignment with the test scales (Czepes, 2009). In order to achieve these objectives, rater training has to be implemented to clarify the qualities of the instrument, to properly address the target test taker population, and to calibrate scoring behavior across and within raters. In fact, training has been found to increase intra-rater consistency (Lumley & McNamara, 1995). However, training does not completely eliminate variance due to rater variability (Lumley & McNamara, 1995; Weigle, 1998).

Numerous studies have been carried out in order to identify and analyze rater effects on the scoring of speaking tests. Some of the factors that have been studied are the effects of rater-ratee interactions, rater main effects (severity or leniency), rater-task type interaction, rater-criteria interaction and gender-based perceptions of behavior (Eckes, 2005).

Some studies have focused on the effects of rater characteristics, in relation to inherent qualities of the raters such as language background, gender, or educational training. However, this type of study has been inconclusive, so that it can be argued that these categorical variables are predisposed to random rather than systematical variance (Brown & McNamara, as cited in Brown, 2010). For instance, Caban (2000) conducted a study to determine whether the differences found between four groups of raters (a total of 83 participants) assessing four

interviews could be attributed to their language background or academic training. The raters had either English or Japanese as their L1, and were from one of four educational background groups (graduate students with EFL or ESL background, ESL teachers, or ESL students). After conducting a Facets analysis to identify possible biases, it was determined that the variation between the four rater categories in this study could not be attributed to language or educational background, even though tendencies of leniency or severity could be observed in the data.

With contrasting results, Winke, Gass, and Mynford (2011) conducted a study in which 107 raters with Chinese, Spanish and Korean as their L2 with various degrees of proficiency rated speech samples from L2 English speakers with those languages as their L1. The researchers reported that "raters with Spanish as an L2 were significantly more lenient toward test takers who had Spanish as an L1, and raters with Chinese as an L2 were significantly more lenient toward test takers who had Chinese as an L1" (p. 3). The researchers concluded that accent familiarity should be addressed in rater training, as it may represent a source of construct-irrelevant variance. Similarly, Kang (2008) studied the relationship between rater's scoring of speech samples and the acoustic measures of *accentedness* of those samples. After performing a Multivariate Analysis of Variance (MANOVA), it was determined that "20 % of variance in proficiency and intelligibility ratings were due to variables relevant to accent (speech rate, pause, and stress), versus variables that are conceptually extraneous (i.e., rater bias)" (p.201).

Such variability or inconsistency of results has also been found with the variable of gender. O'Loughlin (2002) conducted a study in which 16 candidates of the IELTS examination speaking subtest were scored by eight certified IELTS raters, with equivalent number of female and male raters and test taker samples. After performing bias analysis using multi-faceted Rasch measurement, it was found that none of the raters had scored the samples significantly more harshly or leniently towards either gender. It was concluded that "gendered differences are not inevitable in the testing context" (p. 196), and it was noted that other variables may diminish or null the effects of gender interaction effects (O'Loughlin, 2002).

In agreement with O'Loughlin's (2002) findings, the results of a study examining the speaking and writing sections of the Test of German as a Foreign Language (TestDaF) failed to provide evidence of rater effects based on gender (Eckes, 2005). In the case of the speaking subtest, 31 raters examined the samples of 1,348 test takers, each rating between 25 and 134 examinees. After performing a bias analysis and an individual-case analysis, it was determined that the raters did not present any significant pattern of scoring behavior based on gender.

Another factor that has been studied regarding possible unintended effects on scoring is test modality, that is, whether the test is delivered in a direct or semi-direct format. The delivery format of a speaking test and its effects on the assessment process has mainly been studied in relation to their impact on test taker performance in computer-mediated tests. For example, it has been found that test takers' attitude towards the modality of the test is the best predictor of test takers' performance (Yu, 2012). Moreover, while the scores assigned to samples of direct (Oral Performance Interviews) and semi-direct (Simulated Oral Performance Interviews) speaking tests have been found to be comparable, the language functions that can be elicited through either one of the test modalities are rather limited (Alderson & Banerjee, 2002). Given that the comparability of assigned scores for different types of test modality has been a recurring investigation second language assessment research, shouldn't the possible effects of how speech samples are presented to the raters for scoring (e.g., audio vs. audio-visual recording) be studied as well?

Research has been conducted to compare the scoring of audio-recording samples in comparison to the scoring of live performance tests. When only audio is provided, it has been found that the more proficient examinees are affected since their actual level of proficiency is underestimated by the raters. In contrast, examinees with adequate use of nonverbal behavior received higher scores when their performance was video recorded and shown to raters in this format (Nambiar & Goon, 1993). These findings align not only with the fact that higher-ability language learners synchronize speech with nonverbal behavior (Neu, 1990), but also with the point of view from the interactional competence approach to defining speaking ability that nonverbal behavior is, in fact, a part of speaking ability (Ducasse & Brown, 2004).

When it comes to the scoring of recorded speech samples of speaking performance, little has been said about whether the type of recorded speech sample may have an effect on the consistency or severity of rating. Studies of speaking assessment and rater biases choose one of the speech sample types, either audio or video. Nakatsuhara (2007) and O'Sullivan (2002), for instance, made use of videotaped interviews to study interviewee-interviewer effects in the assessment of a speaking test while Ekes (2005) and Winke et al. (2011) made use of audio recordings only. Almost no comparison between the two types and their impact on rating has been made, which is the reason why the current study was conducted.

Although the results are mixed, there have been a number of studies which examined the inclusion of visual stimuli in listening tests (e.g., Wagner, 2007; 2008; Batty, 2015) and its effect on test-taker performance, and "an increasing number of researchers support listening assessments which include as much contextual information in the input as possible" (Ockey, 2010, p.4). Yet, the effect of visual stimuli on raters' performance is an area that remain under-investigated. The only study that compared ratings of audio and audio-visual speech samples is the one conducted by Lavolette (2013), who examined the ratings of audio-only samples, video samples, and samples with audio from the video samples in the context of formative assessment. In their ratings of 39 ESL examinees' performance on the TOEFL iBT direct speaking task, raters were found to significantly favor both types of audio-only samples, contrary to Nambiar and Goon's (1993) findings, so it was determined that the choice of speech sample type could be a factor of unexpected rater variance. Since there might be a difference in the rating process depending on how the speech sample is delivered to the raters for scoring, the main purpose of this study is to explore whether the presence or absence of visual input in the scoring process may have an impact on rater behavior.

## RESEARCH QUESTIONS

The primary purpose of the current study is to determine whether effects of visual input on the rating of speaking test performance can be identified, and whether the mean scores assigned when raters have access to the video recording and when they are scoring with audio recording only are significantly different. The following research questions are addressed.

1. Are there differences in the scoring of a speaking assessment task resulting from the presence or absence of visual input in the speech samples?
2. Do raters report a preference for audio-visual or audio-only recordings of test takers' speech samples when rating? If so, how is such preference justified by the raters?

One hypothesis (H1) is proposed, tested, and discussed in this paper. H1 is based on current findings on rater effects, and presupposes a systematic difference in the perception of test takers' ability after including the visual component in the test takers' speech samples.

H1: The inclusion of visual input in the test takers' speech samples will lead to a significant difference on the scores assigned by the raters.

# METHOD

## Research design

This study follows a quasi-experimental repeated measures single-group design (Wiersma & Jurs, 2009). The experimental variable that was studied (i.e., visual input) was included in the speech samples that were rated in the second scoring session that occurred weeks after the audio-only speech samples were first rated. Mixed methods (Wiersma & Jurs, 2009) were employed to gather and analyze the data since the study involved both quantitative and qualitative analyses of the assigned scores and post-questionnaire responses. Purposeful sampling (Wiersma & Jurs, 2009) was implemented to recruit the raters who participated in this study, given that volunteers were recruited through an online posting on the school website and via email, and that they had to meet certain requirements.

## Context of the study

The study took place at an adult ESL program at Teachers College, Columbia University called the Community Language Program (CLP). The speech samples were obtained from eight students from an upper intermediate proficiency class (i.e., Intermediate 5) when they took the final achievement speaking test. The task that was examined in this study targeted elements of language and speaking skills that were covered during Unit 5, namely, discussing personalities. For example, speaking tasks for Unit 5 involved discussing aspects of personality and how personality may relate to career decisions. For an outline of the course content, see Appendix A. The final number of test taker speech samples used in the study (n=7) was determined after examining the quality of the recordings. Since one of the test-takers interrupted his response repeatedly, seven samples were used instead of the total eight that were originally recorded.

## Participants

Twenty-five raters volunteered to participate in the study. In order to be part of the sample, the raters had to be graduate students of a program in TESOL or applied linguistics, and preferably have had experience in teaching English as a second/foreign language and rating second/foreign language tests. Their ages ranged from 22 to 30, with the mean age of 24.9. Among them, three were male and 22 were female. The raters were heterogeneous in terms of

their teaching and rating experience, but all participants had taught a foreign or second language (24 of them with English teaching experience ranging from six months to ten years). Most of the raters were non-native speakers of English; twelve raters were native Spanish speakers, five were native Chinese speakers, one each of Vietnamese, Japanese, and Korean speakers, and five native English speakers. Half of the participants had already taught at the CLP, or else had had experience teaching in other ESL contexts. None of the raters were told the purpose of this study beforehand. For the analysis of their scores and responses to the questionnaire, each rater was assigned an ID number (from Rater 1 to Rater 25).

## Instruments and materials

### *Speaking test task*

In order to obtain the speech samples to be scored, an achievement test was designed and administered. A blueprint of the test was devised prior to its administration. The unit test was comprised of three elicitation tasks, one monologic and two dialogic tasks, but for the purpose of this study, only one of the tasks was given to the raters for scoring since test taker performance in the dialogic tasks did not reach the desired degree of interactive work, and most semi-direct tests administered on computers use monologic tasks. The elicitation task was taken individually by each participant, and it required them to talk about the types of personality that best fit some occupations (e.g., a teacher, a doctor, a clown). Each test taker was given two minutes to complete the task and was recorded with a videotape recorder. They were asked for their permission for being filmed and tape recorded before the day of the examination. For the language points covered in the unit, test specifications, and the handout that was presented to the students, see Appendices A to C.

### *Rubric*

As part of the achievement test, an analytic rubric was developed in alignment with the goals of the unit being tested (Unit 5 in Appendix A). The construct of speaking ability was operationalized as being comprised of five components: fluency, pronunciation, vocabulary, grammar, and meaningfulness. Each component (i.e., scale) on the rubric ranged from 0 to 4 points. Each point on a scale included a descriptor of the expected performance. Since the task was monologic in nature, a conversational or interactive dimension was not included as part of the measured construct. In addition, given that the unit goals reflected the inclusion of a particular set of vocabulary, this was considered as one of the components of the analytic rubric (for the complete rubric, see Appendix D).

### *Audio and audio-visual speech samples*

After obtaining the speech samples, each was edited so that the raters could focus on scoring a single task (i.e., monologic task), and only the test takers' responses. The audio-visual files were converted to audio-only samples using the AudioDatei video file converter program so that the quality of the audio was equivalent so as not to affect the raters' judgement while

scoring. The instructions as given by the classroom teacher were not part of the speech samples, and they were only given to the raters in written form during the online training session. Based on the suitability for online scoring (e.g. clarity of audio), a total of seven speech samples were selected for the scoring session. Three samples were of male test takers, and four were of females. The test takers were native speakers of Japanese (*n*=2), Spanish (*n*=3), Portuguese (*n*=1), and Arabic (*n*=1), and they had lived in the United States between four months and two years. Three of these samples were used for comparison in this study, which included samples of two Japanese speakers and a Portuguese speaker. Four speech samples were used as calibration exercises as part of norming. Detailed description of the overall procedure and scoring sessions will be discussed shortly.

## *Training materials*

Due to practicality issues, training and scoring were implemented via the online platform *Qualtrics*. Each rater first had to complete a training session, which required them to download and read the rubric and content of the speaking test, go through a presentation of the information, and finally answer multiple questions to confirm their comprehension of the material. Afterwards, raters performed two calibration exercises where they had to listen to a speech sample and rate according to a scale in a matrix table. They were asked to take notes during the calibration exercises so that they could norm themselves after checking their grades that were displayed immediately after submitting their ratings.

## *Scoring materials*

Immediately after completing the online training, raters proceeded to rate three speech samples, also through *Qualtrics*. The format was the same as the calibration exercises, but no official grades were displayed after submitting their ratings. In order to determine the reasoning behind the assignment of scores, raters were asked to provide written comments as they scored each sample.

## *Questionnaire*

At the end of the scoring sessions, raters were asked about their opinions regarding the audio and audio-visual speech samples. Their opinions were categorized, tallied and analyzed. To see the questionnaire, refer to Appendix E.

## *Equipment*

Digital audio and video recorders were used during data collection, and each participant had a computer with internet access to complete the training and rating sessions.

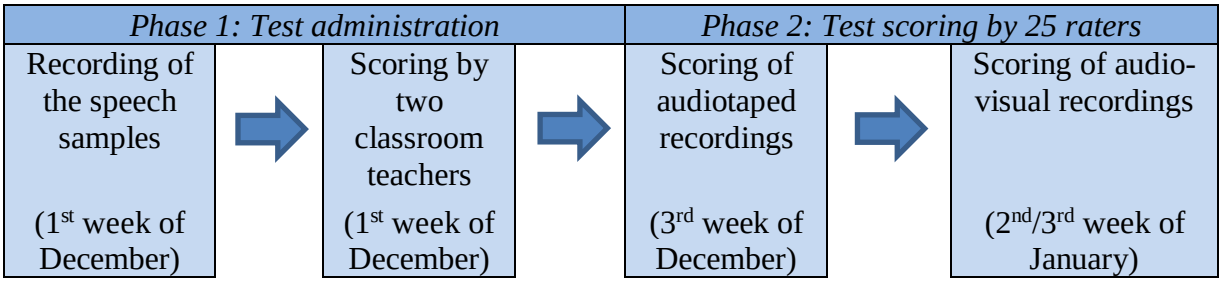
## Administrative Procedures

The achievement test was administered in the final week of the fall term. Each participant's performance was recorded in both audio and video formats to be later scored by the raters. After collecting the samples, the raters were called to participate in the study, which included two scoring sessions. Raters scored the same speech samples from three test takers in both sessions, but first with the audio-only version and three weeks later with the audio-visual version. Four other samples were used for calibration exercises, two for each scoring session. To avoid data contamination, raters did not communicate with each other.

Each online session consisted of a training section in which the context of the program, test takers' level, rating scale and scoring procedures were explained in detail, and a scoring section, in which participants were given the speech samples to be scored. The training section of each session lasted around 20 to 30 minutes. Raters then had to answer questions to verify that they had understood the materials. Afterwards, raters performed two calibration exercises where they had to listen to and rate two speech samples. Using notes they had taken during the calibration exercises, raters were able to norm themselves using the grades they were given immediately after submitting their ratings. After scoring the two practice samples, raters proceeded to score three speech samples used for comparison in this study.

One month after the first scoring session, a second session was held, and raters were asked to score the same set of speech samples, this time with audio-visual input. The same procedures were followed. The calibration exercises included two samples that were different from the ones provided in the first session. The samples used for actual scoring, however, were the same ones that had been previously scored. This was done in order to be able to compare the scores that the raters assigned to these samples with and without the visual input. Since there could be practice effect, the responses were ordered differently to make it less likely for raters to remember the scores they had previously assigned to the samples. Raters were also asked if they recalled having scored the samples before. 12 raters admitted recalling one sample, eight raters recalled two, and five raters recalled all three of them. However, none of the raters remembered the scores that they had previously assigned. The entire procedure is summarized in Figure 1.

**Figure 1.**
**Research Design**



## Data and Scoring Procedures

### *Study variables*

The independent variable in this study is the type of speech sample, which corresponds to either (1) audio-only speech samples or (2) speech samples in audio-visual format. The dependent variable is the average score assigned by the raters to test takers' speech samples as a measure of their speaking ability within the goals of the achievement test.

### *Data coding*

Mean scores were calculated for the average scores assigned by each rater to the three samples that were scored in each session, that is, each of the three samples received two scores by each rater, one when rated with audio only, and one which was assigned when it was presented with audio-visual input. In addition, the justification of each score (as provided in the online scoring platform and the answers to the short questionnaire) were analyzed, first by categorizing them and obtaining the proportions of the different opinions provided, and then qualitatively examining each response.

## Data Analyses

### *Statistical procedures*

In addition to the calculation of descriptive statistics and reliability for the scoring of the speech samples, one additional statistical procedure was conducted. In order to analyze and interpret the mean scores between the two types of input from the samples, a paired samples *t-test* was performed to compare the scores assigned by the raters and identify any significant differences between the two rating sessions (Dörnyei, 2007, p. 215).

### *Computer equipment*

Data was exported from *Qualtrics* to Microsoft Excel 2010. Microsoft Excel was used to organize the data, which was then exported to SPSS Ver. 21.0 to compute descriptive statistics and run a *t*-test.

## RESULTS

## Descriptive Statistics for the speaking test

In order to better understand test taker performance in the speaking task, descriptive statistics were calculated for each of the rating sessions. The five components that comprise the analytic rubric designed for the test of speaking ability (i.e., fluency, pronunciation, grammatical

control, vocabulary, and meaningfulness), were averaged to compute speaking score, which was used to compute the descriptive statistics. In Table 1 we can see the descriptive statistics for both scoring sessions in addition to the scores given by the class teachers.

**Table 1.**
**Descriptive Statistics for Speaking Achievement Test**

| | *Raters* | *N*[1] | *Range* | *Mean* | *Median* | *SD* | *Skewness* | *Kurtosis* |
|---|---|---|---|---|---|---|---|---|
| Classroom teachers' scores | 2 | 3 | .70 | 2.633 | 2.60 | .351 | .423 | - |
| Scoring Session 1 (audio) | 25 | 3 | .49 | 2.386 | 2.272 | .263 | 1.588 | - |
| Scoring Session 2 (audio-visual) | 25 | 3 | .62 | 2.370 | 2.192 | .344 | 1.706 | - |
| Valid N | | 3 | | | | | | |

Regarding measures of central tendency, the mean was 2.63 for the grades that were assigned by the classroom teachers, which is slightly higher than the mean scores assigned by the raters who took part in the study, which were 2.3867 and 2.3707 in the first and second scoring sessions respectively. In addition, the medians were 2.60, 2.272, and 2.192, respectively, which were relatively close to their corresponding means.

The skewness indices were all positive, with values of .423, 1.588, and 1.706, falling within the acceptable range of -2.5 to +2.5. However, this indicates that the test may have been too difficult for achievement purposes. Given that there were only three test takers under comparison, the kurtosis for the test did not fall within the parameters of normal distribution and was not calculable. This is expected of a size this small, but since skewness was within acceptable parameters, the analyses proceeded.

Regarding each speech sample in the first scoring session, as could be seen in Table 2, Sample 1 was assigned a score of 2.20, Sample 2 was assigned a score of 2.688, and Sample 3 was assigned a score of 2.272. In the second scoring session, the average scores fluctuated to 2.192 for Sample 1, 2.768 for Sample 2, and 2.152 for Sample 3. As can be seen in Table 2, no clear tendency could be observed in the small variation between the means, or regarding the change in the degree of variability in the scores as presented by the values of the standard deviation for each speech sample score set. Therefore, in order to determine whether or not systematicity in the differences in the scores could be observed, a paired sample *t*-test was performed. In order to do so, the scores assigned to the three samples by each rater was compared across scoring sessions. This will be explained in detail in the following section.

**Table 2.**
**Scoring results for the experimental scoring sessions (n=25)**

| Speech Sample | Scoring Session 1 | | Scoring Session 2 | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sample 1 | 2.200 | .485 | 2.192 | .416 |
| Sample 2 | 2.688 | .542 | 2.768 | .767 |
| Sample 3 | 2.272 | .382 | 2.152 | .653 |

## Reliability analyses for the test

The data obtained from each scoring session was first used to determine the extent to which raters had systematically assigned scores according to the analytical categories of the speaking achievement test. For this purpose, and in relation to the first research question, the internal-consistency reliability for the first scoring session was calculated and the Cronbach's Alpha was .95 with a standard error of measurement (SEM) of .013. This coefficient shows that 95% of test takers scores can be attributed to their true speaking ability, which is satisfactory.

In the case of the second scoring session, the internal-consistency reliability for the speaking test was calculated as .965. This slightly larger coefficient shows that 96.5% of test takers scores can be attributed to their true speaking ability, which represents a small increase in the consistency of the scoring process. The SEM for this scoring session was .012. Despite this apparent improvement, in order to determine whether or not there was a systematic difference in the mean scores assigned in each session to the test takers, a paired samples *t*-test was performed.

## Comparison of assigned scores across two occasions

The data was organized so that the average scores for each type of speech sample (audio-only or audio-visual) could be compared. In order to determine whether the mean scores assigned by the raters in the audio-only and audio-visual speech sample scoring differed significantly, a paired samples *t*-test was run.

A pair of assumptions had to be met in order to run the *t*-test. First, the sample should be random, and secondly, the variable with the paired differences to be compared should be normally distributed (Weiss, 2010). Even though purposeful sampling was used since raters were volunteers, as long as they met the required qualifications, they were all included in the study as raters, and not selected based on a particular characteristic. As for the satisfaction of normality conditions, when comparing scoring session 1 with scoring session 2, the skewness was of -.151 and .42 respectively, which fall within the parameters of normality ($\pm$2.5). However, it is interesting to see that in the first rating session, there was slight negative skewness, whereas in the second session skewness was positive. They, nonetheless, were close to zero, which means that distribution was close to normal. Kurtosis was established at .239 and -.068, which also fall within the parameters of normality (+/- 2.5). Therefore, the paired samples *t*-test could be performed to compare the scores. Table 3 shows the results of the analysis.

**Table 3.**
**Comparison of the means of each rating session**

| *Pair 1* | *Mean difference* | *Std. Deviation* | *Std. Error mean* | *t* | *df* | *Sig. (2-tailed)* |
|---|---|---|---|---|---|---|
| **RATING1 - RATING2** | .01600400 | .31872788 | .06374558 | .251 | 24 | .804 |

The difference between the means was .016004, while the standard error mean was .06374558. As noted in Table 3, the *t*-statistic of .251 did not surpass the critical *t*-value, and it can be concluded that there is no significant difference in the mean scores that were assigned by the raters across rating sessions, given that the *t*-test failed to show significance at the $p < .05$ level. The paired samples *t*-test result indicates that there is no systematicity in the increments or decrements in the scores that were assigned by the raters ($t = .251$, $df = 24$, $p > .05$, two-tailed). Therefore, it can be determined that the inclusion of visual input in the assessment of the speech samples did not affect the scoring systematically, and the scores assigned in the second rating session fall within the same parameter of the first rating session.

## Results of the questionnaire

After rating the speech samples in the second scoring session, participants were asked to give their opinions regarding the two types of input that had been used for scoring the test takers. In a brief survey, three questions were raised to elicit their perception of the two scoring sessions.

First, they were asked if they had noticed any difference in the delivery quality of the speech samples when they were presented with video as opposed to the audio-only format. This notion was important, given the fact that the quality of the audio was the same in both types of samples, but perceiving this factor as varying across sample types could have led to undesired variation in the scores due to method effects.

Eighteen out of the 25 raters reported that they had noticed a difference between the samples that only included audio and those that also included visual input. From these, 15 reported that they considered that the inclusion of visual input eased the comprehension of the speakers' message, which helped them in the assignment of scores. Among the factors that were mentioned as helpful elements of the videos were body language, facial expressions, and attitudes and feelings (e.g., confusion or awkwardness). Put in the words of Rater 20:

> "The use of video provided a more personal and real evaluation. It allowed me to understand pauses during speech better -when students were examining photos or asking questions to the examiner, for example. Speakers' use of gestures also helped clarify some parts which seemed unintelligible with only the audio."

As stated by this rater, the inclusion of visual input may have allowed for a more comprehensive understanding of the test takers' performance, as it enabled raters to perceive the cognitive processes that were occurring while performing the task.

Of the seven raters who reported that they had not noticed any differences between the samples, three attributed this to not being able to recall the first set of samples, one mentioned that they were exactly the same in quality, one (Rater 10) claimed that she had focused on the audio and had not paid much attention to the video but that "the sound quality of the video samples seems to be better," and two did not explain further. However, the quality and volume of the samples were controlled so that it was the same in both scoring sessions, so the perception of difference in sound quality could only be attributed to the speech samples of the calibration exercises, given they had different speakers as opposed to the speech samples that were scored, which did not change across scoring sessions.

When asked whether they preferred audio-only samples, samples with audio and video, or whether they had no preference, only one of the participants (4%) reported that she preferred audio-only samples, three (12 %) reported that they did not have any preference on the type of input, and the vast majority of 21 (84%) reported that they preferred the speech samples with both audio and video.

The participant who preferred audio-only samples justified her position stating the following: "Audio helps me to focus more on students' speech, whereas video is a little bit distracting" (Rater 10). From her perspective, the inclusion of visual input while scoring was considered as a source of distraction, which made it more difficult for her to focus on the actual performance of the test takers. From this view, factors such as body language and the setting could be regarded as potential sources of bias.

For those three raters who stated that they had no preference regarding the type of input in the samples, one mentioned that the inclusion of video had not affected his judgment during the scoring procedures, while the two others determined that they had focused on the audio to assign the scores to the test takers. Of these, one of them (Rater 6) made an important distinction and stated that "…the video could enhance understanding while the audio eliminates biases." From this perspective, Rater 6 recognizes the video as a potential source of bias, but at the same time acknowledges that it may have a positive effect towards the comprehension of the intended messages of the test takers. This second notion was further explored by those who reported a preference for video speech samples, as discussed below.

From the 21 raters who reported a preference for the video speech samples, 12 mentioned that they preferred the inclusion of video because it incorporated the notion of body language, arguing that this paralinguistic dimension eased the comprehension of the intended message of the test takers (as stated by Rater 5), and facial expressions helped understand attitudes of the speakers. In the words of Rater 12, "body language also provides some reference of the speaker's intentions while trying to convey a message." In this way, it can be noted that these raters consider facial expressions and gestures as enhancing the overall comprehension of the speaker's message, which they consider as an aid in the scoring process. In fact, two raters explicitly mentioned that seeing the test takers' performance was particularly helpful in their role as raters given that they had to provide feedback rather than just listen. In addition, three raters mentioned that the inclusion of video led to a more authentic and natural assessment.

Finally, when asked whether or not they had paid attention to the video, one of the participants (4%) determined that she never focused on the video, five (20%) reported that they had sometimes paid attention to the video, and 19 (76%) stated that they had constantly looked at the video while scoring the samples. This shows that regardless of their opinion towards including video in the scoring process, raters are likely to at least moderately watch the video if it is available.

# DISCUSSION

The first research question of this study enquired whether or not differences in the consistency of speaking assessment scoring would be found depending on the presence or absence of visual input. It was found that the inclusion of visual input slightly increased the internal-consistency reliability, from .95 to .97 using Cronbach's Alpha. This improvement in the test reliability shows an increase in the systematicity of the scoring process, yet it does not explain whether the scores assigned to the test takers would remain equivalent. Furthermore, this slight increase in internal consistency could be due to the repetition of the training and scoring procedures. The mean scores of the three test taker samples in each session were compared using a paired *t*-test analysis, and the result shows that the means, although slightly different in their values, are not statistically different suggesting that the assigned scores remained within the same range. Therefore, Hypothesis 1, which was developed in relation to this research question, could not be proven since it expected that the inclusion of visual input in the samples would have a significant effect on the scores assigned by the raters. Given these results, it can be hypothesized that the inclusion of visual input did not have a systematic impact on the assignment of scores.

Qualitatively, however, it was observed that raters did have a clear opinion in relation to the type of speech sample used. The second research question asked whether or not raters would report a preference for audio-visual or audio-only speech samples. Results showed 76% of the raters admitted a preference for the inclusion of visual input in the speech samples to be rated. In doing so, the major reason that raters provided to justify this preference was that body language and facial expressions allow for a more authentic experience, so that the intended message of the speaker and the delivery of the speech is better understood (e.g., noticing the reasons for pausing, attitudes towards the interlocutor and interaction with the materials). Similarly, 84% of the raters reported having constantly paid attention to the video in contrast to the 4% who claimed to have ignored the visual input during the second scoring session.

These results, while limited in generalizability given the small sample size and inclusion of a single task, shed light on a topic that has not yet been studied in depth from a neutral perspective. The majority of raters who participated in this study advocated for the inclusion of visual input as a way to complement and ease the understanding of test taker message and performance, thus enhancing the accurate assignment of scores. While such perception did not have a significant effect in the differences in means across scoring sessions, it did have a positive effect on the raters' satisfaction with their own performance as raters. Therefore, it may be argued that, even if the inclusion of visual input does not systematically affect the assignment of scores, it may be implemented as an important way of authenticating the communicative dimension including non-verbal features that are involved in the scoring of speaking assessments. This may result impractical in some contexts, so that it may not be readily applicable unless there is a rationale for enhancing perceived ease of rating. The findings of this research can also call for a research agenda that further explores how test scoring method can have an effect on rating, and that it provides some insights on the importance of examining rating processes and procedures that depart from traditional views.

## CONCLUSION

To conclude, in examining the possible effects of the inclusion of visual input in speech samples for the scoring of an achievement speaking test, no significant differences in the means of the scores assigned to the test takers were found when a paired *t*-test was performed. This means that in spite of the observed variability between each sessions' scores, these remain within a comparable range, and given the high internal-consistency reliability attained in both occasions, it can be determined that visual input cannot be regarded as a systematic source of variation in the scoring of these speech samples. Nevertheless, there was a clear tendency among raters to opt for the inclusion of visual input in the speech samples when scoring because it allowed for a more complete and straightforward scoring experience. More research needs to be done on this issue, and a series of recommendations are provided below.

## LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

One limitation of this study is that the sample size was small, and the findings could only be generalizable to similar contexts. In addition, other independent variables which may be sources of bias such as L1, gender or rating experience are not integrated into the study, which affects the validity of the design.

Another limitation is the time span in between each scoring session. More time should be placed between scoring sessions, and distractor samples should have been used since even though no specific scores were recalled from the raters, the use of the same speech samples could have had an impact on the reliability of the scoring due to practice and method effects.

As a matter of fact, there are a couple of caveats to the online training and scoring system. First, it can be argued that the training was too short when compared to a regular training program. Similarly, the number of speech samples to be scored was small due to the availability of the raters and other resources. Regarding the time raters took to complete each rating session, *Qualtrics* revealed that each rater spent a different amount of time to complete the training and the scoring program. While no absolute assumption can be made, more control regarding time spent on the training and scoring should be implemented. It should also be noted that the large number of raters may have helped attain such favorable values in the reliability analyses.

Finally, only one type of test task (i.e., monologic task) was analyzed in the study and the rating scale was restricted to the language components that could be evaluated through this particular type of task, so other dimensions of oral language (such as conversational ability or pragmatic control) were not incorporated into the construct of speaking ability upon which the rubric was designed. The incorporation of a dialogic task could have yielded different results, since nonverbal behavior has been found to be an important element in face-to-face oral communication.

Increasing the number of raters and speech samples to be scored would highly benefit the next version of a similar study. Also, standardization of the timing for each training session should be carefully controlled for. Moreover, the inclusion of other types of tasks should be explored, so that other variables (e.g., task complexity, interaction, length of the expected response, proficiency levels, and test takers characteristics) could also be explored. In addition, a

deeper analysis, which could include bias analysis, should be implemented to make sure that the inclusion of visual input does not trigger sources of bias.

## REFERENCES

Alderson, J. C. & Banerjee, J. (2002). State of the art review: language testing and assessment (part two). *Language Teaching, 35*(2), 79-113.

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

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher and F. Davidson (Eds.). *The Routledge Handbook of Language Testing.* New York, NY: Routledge.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), Pp. 1-44.

Carr, N. (2011). *Designing and Analyzing Language Tests.* Oxford, UK: Oxford University Press.

Csépes, I. (2009). *Measuring Oral Proficiency Through Paired-Task Performance*. Berlin, Germany: Lang.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics.* Oxford, UK: Oxford University Press.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing, 26*(3), 423–443.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221

Fulcher, G. (2003). *Testing second language speaking*. London, UK: Longman.

Ginther, A. (2013) Assessment of speaking. In C. A Chapelle, (Ed.). *The Encyclopedia of Applied Linguistics.* Oxford, UK: Wiley-Blackwell.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative Impact of Rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* Volume 6, p. 181–205.

Lavolette, E. (2013). Effects of technology modes on ratings of learner recordings. *IALLT Journal of Language Learning Technologies, 43* (2) 2013.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71.

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

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal, 24*(1), 15–31.

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. Scarcella, E. Andersen, & S. D. Krashen (Eds.), *Developing Communicative Competence in a Second Language* (pp. 121–138). New York, NY: Newbury House.

Nakatsuhara, F. (2006). Impact of inter-interviewer variation on analytical rating scores and discourse in oral interview tests. *Newcastle Working Paper in Linguistics, 12*, p. 55-68.

Ockey, G. J. (2013). Assessment of listening. In C. A. Chapelle (Ed), *The Encyclopedia of Applied Linguistics*. USA: Blackwell Publishing Ltd.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

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

O'Sullivan, B. (2014). Assessing speaking. In Kunnan, A. J. (Ed.) *The Companion to Language Assessment, First Edition*. UK: John Wiley & Sons, Inc.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(1), 67–86.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly, 5*(3), 218–243.

Weigle, S. (1998). Using FACETS to model rater training. *Language Testing, 15*(2) 263–287.

Wiersma, W. & Jurs, S.G. (2009). *Research Methods in Rducation. An Introduction*. Boston, MA: Allyn & Bacon.

# APPENDIX A

## Intermediate 5 Contents

| Unit/Theme/ | Grammar & Vocabulary | Listening and Speaking | Reading and Writing | Tasks/ Projects |
|---|---|---|---|---|
| Unit 5<br>The Real You?<br>Personalities<br><br>Sept 25-Oct 9 | **Gerunds as**<br>**-subjects**<br>**-objects of verbs**<br>**-objects of preposition**<br>**Verbs followed by gerunds and infinitives**<br>**Vocabulary on Personality** | **Listening:**<br>**-Taking notes**<br>**-Listening for detail**<br>**-Listening for gist**<br>**Pronunciation**<br>**-Reducing of**<br>**-Speaking**<br>**-Discussing about feelings.**<br>**-Discussing about personality.** | **Reading:**<br>**-Use of graphics.**<br><br>**Writing:**<br>**-A Personal Letter**<br>**-Distinctions between formal/informal letters** | **-Writing your horoscope: Students will put into practice their knowledge of vocabulary related to personalities and the use of gerunds and infinitives.** |
| Unit Objective | At the end of the unit students will be able to communicate about personalities and will have learned to use the appropriate object of verbs. | | | |
| Unit 6<br>If I Had My Way<br>Wishes<br>Oct 10-Oct 24<br><br>Additional theme:<br>Human Rights | -The Second Conditional<br>-Asking for and giving advice | Listening:<br>- Listening to summarize -<br>Pronunciation<br>Rhythm<br>-Speaking<br>-Encouraging and discouraging | Reading:<br>-Guessing vocabulary from the context.<br><br>Writing:<br>-Analysis | -Ask Alice<br>Students will host a support radio show and will write an advice column. |
| Unit Objective | At the end of the unit students will be able to communicate their ideas about unreal situations, as well as support or discourage a position in written and spoken forms. | | | |
| Unit 7<br>What's so funny?<br>Humor<br>Oct 28- Nov 7<br><br>Holiday theme:<br>Halloween<br><br>Additional theme:<br>Slang and humor. | -Reported Speech<br>-Backshift<br>-Reporting verbs | Listening:<br>-Listening for definitions.<br>-Using Stress to check understanding.<br>-Speaking<br>-Reporting someone else's thoughts, ideas and comments. | Reading:<br>-Tone Recognition<br>-Making inferences from the passage.<br><br>Writing:<br>-Definition paragraphs.<br>-Short Stories. | -A Ghost story: to practice reported speech within the holiday spirit, students will write a ghost story to share with the class. |
| Unit Objective | At the end of the unit students will be able to report back statements and ideas, recognizing when there is a shift in tone in the verb. | | | |
| Unit 8<br>So That's How…!<br>Processes | Describing a Process | Listening:<br>-Identifying Steps in a process. | Reading:<br>-Chronology in processes | -Presentation.<br>Students will prepare a |

| | | | | |
|---|---|---|---|---|
| Nov 11-Dec 4<br><br>Holiday theme: Thanksgiving<br><br>Additional theme: Wine | -The Passive Voice<br>-Using an Agent<br>-Sequential linkers | Pronunciation<br>Stressing new information.<br><br>Speaking<br>-Sequencing a process.<br>-Expressing interest or indifference. | Writing:<br>-A Process paragraph<br>-A recipe | presentation to explain the process that is followed to make a product. |
| Unit Objective | At the end of the unit students will be able to develop a chronologically sequenced process in both written and spoken forms. | | | |

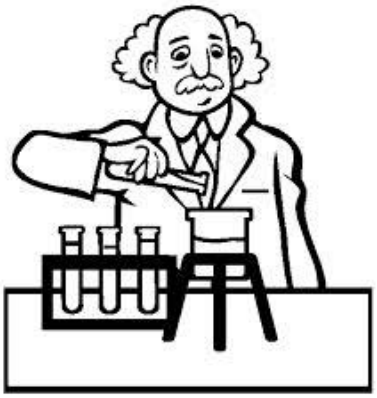
# APPENDIX B

## Test task specifications (adapted from Bachman and Palmer, 1996)

| | *Task 1* |
|---|---|
| **INPUT** | |
| **Format** | |
| Channel | Aural and visual |
| Form | Language and pictures |
| Length | Short (discourse) |
| Type | Extended-production |
| Vehicle | Live/ Reproduced |
| **Language Characteristics** | |
| Organizational characteristics | |
| Grammatical | |
| Textual | Variety of forms |
| Pragmatic characteristics | Variety of forms |
| Functional | Variety of forms |
| Sociolinguistic | Variety of forms |
| **Topical characteristics** | Personalities |
| EXPECTED RESPONSE | |
| **Format** | |
| Channel | Aural |
| Form | Language |
| Length | Extended (Essay) |
| Type | Extended-production |
| **Language Characteristics** | |
| Organizational characteristics | |
| Grammatical | Variety of forms |
| Textual | Variety of forms |
| Pragmatic characteristics | Variety of forms |
| Functional | Variety of forms |
| Sociolinguistic | |
| **Topical characteristics** | Personalities |
| Reactivity | Non-reciprocal |
| Scope of relationship | Broad |
| Directness of relationship | Direct |

# APPENDIX C

Below are some pictures of a few jobs.
Is personality an important factor for a job?
Discuss the types of personalities that you think are more suitable for some of the following jobs.
You will have 2 minutes to discuss and agree on something.



# APPENDIX D

## Rubric for the Speaking Test

| | |
|---|---|
| **Grammar** | |
| 4 | The response is grammatically accurate and it displays complex structures. Only minor errors are made. |
| 3 | The grammatical structures being used are for the most part accurate and do not impede communication. The response may use somewhat simple structures or include frequent minor errors. |
| 2 | There is some control over the grammatical structures, but errors are frequent and impede comprehension of the message. |
| 1 | Grammatical errors are too frequent, systematic, and often impede communication. |
| 0 | There is not enough evidence of any type of language control. |
| **Vocabulary** | |
| 4 | The response provides a wide range of vocabulary. The adjectives used to describe personalities were used appropriately with regards to their meaning or form. |
| 3 | A good range of vocabulary was employed by the test taker. The adjectives used to describe personalities were appropriately used, even though they may belong to the target vocabulary and instead came from high frequency vocabulary. |
| 2 | Attempts to use less frequent adjectives for personality were partly successful, but errors with the form of the adjective were made. |
| 1 | The vocabulary was of very high frequency, errors were made with the form and the words may have been repeated or description was used rather than the exact word. |
| 0 | There is not enough evidence of any type of language control. |
| **Pronunciation** | |
| 4 | Pronunciation is accurate and intonation and stress patterns allow full intelligibility. Infrequent pronunciation errors are made but do not impede intelligibility. |
| 3 | Speech is mostly intelligible, with only infrequent errors that seldom affect intelligibility. Phonemic accuracy, intonation and stress are good and it does not affect the meaning. |
| 2 | There are problems of intelligibility. Systematic phonemic errors hinder comprehension; mistakes in stress and intonation make comprehension more difficult. |
| 1 | The message is often misunderstood and communication repairs are frequently done due to problems in pronunciation. Unintelligibility of speech affects comprehensibility. |
| 0 | There is not enough evidence of any type of language control. |
| **Meaningfulness** | |

| | |
|---|---|
| 4 | The message being conveyed can be fully understood despite any type of errors. |
| 3 | Most of the message can be conveyed without difficulty, errors do not obscure the meaning even if they demand careful attention from the listener. |
| 2 | It is difficult to comprehend the intended message of the speaker. |
| 1 | Only part of the intended message can be understood. |
| 0 | There is not enough evidence of any type of language control. |
| **Fluency** | |
| 4 | Speech is naturally fluent, the use of pauses resembles authentic speech and repairs are made effectively. |
| 3 | The response is fluent, pauses and restarts are used so that they do not impede communication. |
| 2 | The speaker makes use of long or frequent pauses, which in addition to restart, may affect the comprehension of the message at hand. |
| 1 | Speech is too fragmented; unnatural pauses and overuse of restatements hinder intelligibility. |
| 0 | There is not enough evidence of any type of language control. |

# APPENDIX E

## Questionnaire (Administered through Qualtrics)

1. Did you notice any difference between scoring the samples that included scoring audio only and those that included video?
   Explain.

2. Which of the types of input do you prefer?
   a) Video
   b) Audio-only
   c) I have no preference.
      2.1 Why?

3. Did you pay attention to the video?
   a) Yes, constantly.
   b) Yes, sometimes.
   c) No, I ignored it.