An Investigation of Interrater Reliability in Oral Testing

Gregory Hadley and Jerry Mort
An Investigation of Interrater Reliability in Oral Testing

Gregory Hadley and Jerry Mort

Nagaoka National College of Technology
September 6, 1999

Abstract

This report examines the interrater reliability of oral communication tests at Nagaoka National College of Technology. The procedure, research instrument, analysis, results and the measures that need to be taken in order to bolster reliability will be discussed. The findings of this report suggest that considerable negotiation between examiners is necessary before highly significant levels of interrater reliability can be reached.

Introduction

An emphasis upon the testing of spoken English as a foreign language has steadily increased over the past several years in Japanese Colleges and Universities.1 Because of this trend, a number of methods for testing students in their oral communicative ability have been included in the classroom. One important issue which needs consideration when creating oral tests is that of interrater reliability. Interrater reliability, as defined by Weir, measures the consistency of scoring between different raters.2 As it relates to this study, interrater reliability is the degree of correlation between two or more examiners, with the goal of determining whether they are using the same set of criteria when testing the oral proficiency of their learners. If the level of reliability between two or more examiners reaches the level of significance, this may indicate that students in different classes are receiving fair treatment in terms of testing. This paper, which should be considered a "work-in-progress", examines the interrater reliability of the oral communication tests at Nagaoka National College of Technology. The procedure, analysis, results and the measures that need to be taken in order to bolster reliability and validity will also be discussed.

Method

In this section, we will discuss the subjects, materials used in the testing of the subjects, and the procedures taken to insure interrater reliability. This study was conducted by correlating and comparing the scoring of the two authors, who are presently engaged in the administration of oral tests at Nagaoka National College of Technology. The class chosen to participate in this research was a high school level Oral Communication course designed to encourage fluency in the target language. The subjects were also chosen because their class has

Received September 3, 1999.
*Department of General Education Nagaoka National College of Technology)
been split into two equal groups (Groups A and B) due to the large number of students in the class this year. This is not a normal occurrence at this school, and there has concern among teachers about how to assure that both groups of students are receiving the same level of instruction in all their classes. The authors are presently the only members of the teaching staff engaged in oral testing, and represented the total population of examiners for this study.

Subjects

The subjects chosen to participate in the oral testing segment of this study were Group A of the second year Japanese students (n = 28) in the Materials Engineering Department at Nagaoka National College of Technology. Most of the subjects were raised and educated in cities and villages within Niigata Prefecture. This group was considered homogenous in terms of age, mother tongue, academic status and cultural background. Most of the learners appear to be false beginners in terms of oral proficiency. The two examiners themselves formed a small but a very homogenous group in terms of age, educational level, language, social upbringing, race, and sex.

Figure 1

Speaking Test Evaluation Sheet

<table>
<thead>
<tr>
<th>Name ____________________________</th>
<th>Student Number ____________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class ____________________________</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale</th>
<th>0 • 1 • 2 • 3 • 4 • 5</th>
</tr>
</thead>
</table>

| a | Communicative Ability | Includes length of utterance, flexibility to speakers of differing levels. Complexity of responses. (x 6) |
|---------------------------------|---------------------------------------------------------------------------------------|
| b | Fluency | Appropriate Speed, Pauses, and Discourse Strategies. (x 4) |
| c | Vocabulary | Did the student use a wide variety of words and phrases, or use new vocabulary learned in class? (x 3) |
| d | Nonverbal Strategies | Did the student supplement oral communication with appropriate gestures, eye contact and body language? (x 3) |
| e | Grammar | How accurate and appropriate was the student's grammar? (x 2) |
| f | Pronunciation | Was effort made to use correct intonation, or was the accent a barrier to communication? (x 2) |

<table>
<thead>
<tr>
<th>Scoring Guid</th>
<th>0 - 15: Repeatedly communicats in first language or uses first language strategies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>16% - 29%</td>
<td>Unintelligible and incoherent.</td>
</tr>
<tr>
<td>30% - 49%</td>
<td>Mostly unintelligible. However, a native speaker might understand some of what learner is saying, but only after considerable effort.</td>
</tr>
<tr>
<td>50% - 59%</td>
<td>Somewhat understandable. Discourse seems non-communicative, and memorized only for the examination.</td>
</tr>
<tr>
<td>60% - 69%</td>
<td>Learner is making a sincere attempt to communicate. Although discourse causes cognitive strain for a native speaker, he/she can understand most of what is said.</td>
</tr>
<tr>
<td>70% - 79%</td>
<td>Good communicative ability. Fairly fluent, uses new vocabulary, grammar &amp; pronunciation adequately.</td>
</tr>
<tr>
<td>80% - 89%</td>
<td>Superior communicative ability and fluency. Uses new vocabulary, mostly accurate grammar &amp; pronunciation, good use of non-verbal strategies.</td>
</tr>
<tr>
<td>90% - 100%</td>
<td>Excellent communicative ability. Language skills are no longer a barrier to communication.</td>
</tr>
</tbody>
</table>

Figure 1
An Investigation of Interrater Reliability in Oral Testing

Materials

The text used for this course was *Communicate.* From our experience with the learners at this school, this text is well matched to their level of oral communicative ability. Test questions were taken directly from the textbook, and were made in the form of cue cards that students would draw at random and ask others in their group. The tool used for measuring the students' oral proficiency was developed in-house by the first author. It represents a composite of elements from the Cambridge PET and KET testing tools, testing instruments used at the Kanda University of International Studies, and the suggestions for speaking tests in the book by Cohen. An example of the testing sheet can be seen in Figure 1.

The internal reliability of the testing instrument was investigated using the split-half method. The odd rating bands were divided from the even-numbered rating bands, and then correlated using the Pearson $r$ correlation coefficient. The reliability for this test came to +0.54. This would be an unacceptable level of internal reliability for an objective test such as a multiple-choice test. However, Heaton explains that lower levels of internal reliability in instruments that measure oral communicative ability are not only inevitable (due to the numerous uncontrolled variables in both the students and the environment), but a moderate level of internal reliability is actually preferable. For an oral test to have a high level of internal reliability, it would need discreet, direct and objective questions that would require equally discreet answers, which then could be determined to be either 100% correct or incorrect. Questions such as these would likely disrupt opportunities for authentic communication on a speaking test and thus destroy the test's validity. Therefore, with Heaton, we feel that a superior instrument that measures oral proficiency should be moderately reliable and highly valid. Dependent on whether one accepts the split-half method as a truly valid method for determining the reliability of this test, we believe that the materials used in this study may be sufficiently reliable for our purposes at Nagaoka National College of Technology. An internal reliability measure of +0.54 suggests that the test may be flexible enough to allow true communicative expression to take place. We also feel the test contains a level of face validity, mainly because it was designed and weighted to measure aspects of oral communicative ability regularly encouraged in class. To strengthen the claim for content validity, all of the questions were taken from textbook activities that the students participated in while in class, and represented a fair sample of what was presented and taught in the course.

Procedures

Before the beginning of the first semester in spring, 1999, the second author was trained in the procedure for administering this speaking test. Three fourth year students at the Nagaoka National College of Technology posed as practice subjects. The procedure was taught as follows:

- Groups of three students enter the room and take their seats.
- The examiner asks the students to choose two cue cards, which contain discussion questions from the textbook.
- Each student asks one of the other people a question from the two question cards that he or she has chosen. The student who is questioned then reflects the same question back (*how about you?* or *What do you think?*).
- After answering, the student then asks the second question card to the other person in the group, who answers and again reflects the question back.
- This process is continued by the two other students, so that each person has the possibility of up to six different opportunities to speak and respond to another in English.
- However, the time limit is normally limited to five minutes a group, so depending upon the fluency of the members, often less than six verbal exchanges take place.

The second author went through this process...
two times with the practice subjects, and graded them using the evaluation instrument in Figure 1. The first author also graded the practice subjects. Differing scores were then discussed, as well as possible ways to make our rating more consistent with each other.

During the first semester, in preparation for the speaking test, the second author (who was the classroom teacher for Group A) prepared the students by verbally explaining to them in English about the requirements of the test. He also led the entire class to form groups and participate in a series of practice tests the week before. On the day of the actual test, the second author conducted the tests and rated the learners. The first author sat off at a distance and rated the learners as well. Testing the entire group of learners took approximately 50 minutes. To avoid contaminating the data, the scores of the students were not discussed between us until after all the students had taken the test and the grades of each student had been individually determined.

**Results**

The data from both examiners was collated and compared using the teachers' software package *VAR Grade for Windows*. Because there were only two examiners, the method of analysis was set up as a directional one-tailed measure, which used the Pearson \( r \) correlation coefficient. With a probability of only one percent that the findings were by chance alone \( (p < 0.01) \), the critical level of significance for test scores of a group of 28 subjects is at approximately +0.46. The interrater reliability in this study reached +0.47, just barely allowing us to reject the null hypothesis (see Figure 2).

While the level of interrater reliability is significant, we can see other problems, such as the potential for variation and errors, as seen in the spread of the scattergraph in Figure 2. Other data also reveals this variability. The following graphs show the distribution of scores. Figure 3 shows the distribution of scores for Examiner 1. In Figure 3, the mean is 65.14%; the median is 65%, and the standard deviation of 6.39. Figure 4 shows the distribution for Examiner 2, with the mean at 59.5%, the median at 62.5%, and the standard deviation at 9.8. Though the tests were statistically headed in the right direction in terms of validity, in actuality, if the first author had...

![Figure 2](image-url)
been the classroom teacher, a greater number of students would not have passed the test. These findings also beg the question of whether or not the other group of second year students taught by the first author, Group B, might be measured by a stricter standard. It is likely that their grades would be higher, were they in Group A.

Discussion

We were surprised and quietly disappointed with these uninspiring findings. All of the data suggested that we still had a lot to talk about and negotiate before being able to increase the level of reliability in our grading. In discussing the possible causes for these results, we came to the following conclusions:
Despite earlier practice, the first time nature of the actual testing session caused the second author to experience a slight lack of confidence and feelings of distraction.

The subjective nature of an oral communication test may have had a part in the wide variance of scores. The scoring bands and their meaning need to be made more explicit.

Related to this point, it was discovered that the first author used a basic criterion of "would a Native Speaker, (such as the type living in the Midwestern part of the United States), who is unaccustomed with Japanese speech patterns and mannerisms, be able to understand this student?" The second author's approach was more akin to, "Based upon my experience of living in Japan for eight years, can I understand what this student is trying to say?"

We concluded that, despite the attempt at training in the procedure of testing and discussing ways to make the test scores similar, we were not explicit in our basic pedagogic criteria for rating students. After some debate on the pros and cons of both criteria, we worked out a middle ground between the two, which was stated as, "will a native speaker of English, who is sincerely open to communicating with Japanese, be able to understand what the student is trying to say, even though he or she is mostly unaccustomed with Japanese mannerisms and speech patterns?"

As stated earlier, it was discovered that the scoring bands are not as explicit as previously thought. In the past, this testing instrument was used only by its creator, who clearly understood the meaning of the scoring bands. As more people begin to use this instrument, the difference between a student who makes a "5" as opposed to a "3" will need to made clearer. With relation to the use of the scoring bands, we learned that at the beginning of the test, one examiner began rating all of the learners automatically at 60%, (usually "3" on the scale, see again Figure 1), and then moving either up or down the scale based upon their performance in the test. The other examiner went on first impressions, whether high or low, and moved up or down according to the student's overall test performance. It was decided that on future tests, we would both start at 60%, and adjust the grade accordingly.

We also identified a number of weaknesses in the study itself. One was the method of determining the test's internal reliability. Although the split-half method is used with success with many more objective test designs, it is not certain if our test instrument can be measured objectively. We suspect that this instrument is more organic in nature, and cannot be easily separated into different parts. If time allows, test-retest will be used to determine the internal reliability of this test. This may provide a clearer and fairer measure of the test's internal reliability as a means of measuring oral communicative ability.

Yet, even if these conditions are satisfied, there is also the problem of only having two examiners for testing interrater reliability. It would be better if we could have had more examiners involved in this study. More examiners would likely increase the validity of this research.

Nevertheless, rather than seeing this study as a failure, we are encouraged that the preliminary results of our research have helped us to come to some important decisions about what to do for the next battery of oral communication tests. We believe that these decisions, coupled with a better understanding of the other's pedagogical stance, will help to bolster the level of interrater reliability. Future research will seek to determine the reliability of the speaking test itself, and also seek to ascertain if the added insight gained from the discussions between the examiners actually has increased the level of interrater reliability. If so, as stated earlier in this paper, stronger claims to validity might then be made, and perhaps then this oral testing methodology can be confidently applied to other EFL classes within the school.
References


