A Study of Interrater Reliability in an In-house Oral Proficiency Test

Soo-im Lee

オーラル・プロフィシェンシィ・テストにおいての信頼性・妥当性に関しての研究

李　　洙　　任

Abstract

Many oral interview tests such as the Foreign Service Institute (FSI), the Speaking Proficiency English Assessment Kit (SPEAK), the Oral Interview (OI) and the Oral Proficiency Interview (OPI) are used in the U.S.A. However very few standardized tests of oral skills are available in Japan. Most oral proficiency tests used are in-house tests designed by individual institutes and the validity and reliability of those tests are somewhat questionable. This study attempts to measure the validity and reliability of an in-house oral proficiency test by focusing on the interrater reliability. It was found that there were multidimensional causes for the judges' subjectivity in their evaluation process. The findings in this study are closely linked to the variables vital to solving the problems with the interrater reliability of the test and also enabling us to find out which attributes are important to measure in our students' oral performance.

Key Words: Communicative Competence, Oral Testing, Language Measurement
(Received September 6, 1996)

抄　　録

米国においては、the Foreign Service Institute (FSI), the Speaking Proficiency English Assessment Kit (SPEAK), the Oral Interview (OI), the Oral Proficiency Interview (OPI) など、スタンダード化された多くのオーラルインタビューテストがあるが、日本においてのオーラルスキルをテストする試験は、公式テストにおいては実用英検、国連英検のみで、それ以外は個々の学内又は組織内テストである。目的に応じてテストが適切かつ正確であるかを判断する重要な要素として信頼性、妥当性があるが、現在日本で実施されているオーラルテストは、高い信頼性、妥当性を達成し得ているであろうか。本研究では、ある組織内テストでの採点者のテスト実施手順や評価基準においての相違に焦点を置き、それがどのように評価結果に反映するかを研究課題とする。採点者間の一律性を高めることがテストの質を高めることとなり、より信頼性、妥当性の高いテストに向けての今後の課題、示唆を述べる。

キーワード：コミュニカティブ・コンピテンス、オーラル・テスティング、
ラングェッジ・メジャメント
(1996年9月6日　受理)

1. Introduction

One of the most important areas all language teachers should be concerned with is how to find accurate measurements of their students' abilities. Vast quantities of resources are available about testing; however, designing adequate tests is not an easy task for teachers. The misuse of language tests is minimized when test users obtain two sources of information as their knowledge base. One is a basic knowledge of the testing principles of validity and reliability, which is available in a number of publications from various sources. The second source of knowledge is a firm understanding of the features and quality of the test that is currently in use. Although communicative language testing has been regarded as important in TESL in recent years, communicative competence is a complex unit to be defined clearly and needs multidimensional observations to be tested.

Designing reliable oral tests is an especially difficult task. The difficulty is derived from not only the practical difficulties of oral testing, but also the difficulty of achieving a high level of validity and reliability in the tests. Why are oral tests so difficult to design and often inaccurate? There are two main sources of inaccuracy in oral tests. The first source concerns what attributes we want to measure in oral testing, but the question often remains obscure. If the objective of teaching spoken language is the development of the ability to interact successfully in that language, test users should pay attention to the validity of the test and assess whether the tests they use are adequately designed to measure those skills. We want to set tasks that form a representative sample of the population of oral tasks that we expect candidates to be able to perform. The tasks should elicit behavior which truly represents the candidates' ability and which can be scored in terms of validity (Hughes, 1989). The second source concerns the reliability in scoring the oral test. This is the basic problem in testing oral ability. Holistic scoring (often referred to as impressionistic scoring) involves the assignment of a single score to a piece of oral performance on the basis of an overall impression of it (Hughes, 1989).

Since the oral test is usually conducted in a face to face interview test situation, the subjectivity of the interviewer is reflected in both areas of procedure and evaluation. The scoring, particularly in interview tests, tends to be more impressionistic than scoring in other types of tests. When a degree of judgment is called for on the part of the scorer, as in the scoring of the performance in an interview, perfect consistency is not expected. The purpose of this study is to identify possible and important variables in this phenomenon of the judges' subjectivity and consequently to find clues as to how we can make oral tests as valid and reliable as possible. There

are few empirical studies on interrater reliability of oral testing. Therefore, it is hoped that this study will contribute to the establishment of a new research niche.

2. Review of the Literature

Certain authors have suggested how high a reliability coefficient we should expect for different types of language tests. Lado (1961), for example, says that good vocabulary, structure and reading tests are usually in the .90 to .99 range, while auditory comprehension tests are in .79 range. He adds that a reliability of .85 might be considered high for an oral production test but low for a reading test. Where does this compromising view come from and also why does the subjectivity level increase in oral testing compared to the other testing areas such as reading or vocabulary tests. In spite of the difficulties in designing oral tests, many efforts have been made over the past decade to develop and refine tests of productive language ability, including tests of oral communicative proficiency (Bachman & Palmer, 1982). Such efforts are very important and also significant for all test users in order to avoid harmful backwash effects in learning and teaching. For example, Jafarpur (1988) found in his study on FSI that the average of three judges' ratings is a better appraisal of the testees' true ability than that of any single judge's or pair of judges' ratings. The study indicates that there are a great deal of discrepancies among the judges' ratings so that the averaged score of all judges are more reliable than the individual judge's score. Also he found in his experimental study, using multiple regression, that any two of the five components which were used for FSI, (grammar, pronunciation, vocabulary, fluency, and comprehension) may correctly predict the oral proficiency of the testees. This finding is useful for test administrators to know especially when the judges of oral testings have to give the tests and evaluate the testees at the same time. Falk (1984) also questioned whether grammatical correctness and communication can be tested simultaneously. She states that in the oral tests we continue to maintain that effective communication is the main criterion for success. We do not, however, have a clear or specific, and at the same time manageable, definition of what this is. We are able to count errors, but we cannot quantify communication. Our in-house oral proficiency test also provides five components to be used in the marking scheme, and these five components are: word power, grammar, pronunciation, comprehension, and fluency. In this study the effectiveness of providing these components as marking criteria to increase interrater reliability of the test is also discussed.

3. Research Design and Methodology

In quantitative research, the aim is to gather objective data by controlling human and other extraneous variables and thus gain what they consider to be reliable, hard data and replicable findings. However, the questions for this study particularly focus on the mental mode of the judges and an in depth study of the test –giving process so a true experimental research design is not suitable for the aims of this research. It was believed that strict methodological study was not adequate because a hypothetic deductive paradigm to make a generalization would set a constraint in this study. A case study approach was chosen for this research for the following reasons. A case study researcher focuses on a single entity, usually as it exists in its naturally occurring environment (Johnson, 1992). Case study methodology is flexible and is formulated to suit the purpose of this study. The goal of this study is to provide a "descriptive and interpretive–explanatory account of what the interviewers do in our oral testing situations." Naturally occurring data was collected for this study without manipulation and by a triangulation strategy of multiple resource data, direct observation and observation of the video–taped records, and interviews with the judges. No controlled experimental data were collected; however, two pieces of quantitative of statistical information were included. These were, namely the coefficient rate of the test to measure reliability of the test using the test – retest method of estimating internal–consistency reliability and the standard error of measurement of the test to see the differences between the actual scores and the true scores of this oral test. This study was conducted to provide a clearer picture of the judges' evaluation processes in our oral proficiency test, and if significant features are collected in the data, the findings might contribute to giving clearer operational definitions to the significant variables for future studies. The data collection was completed over two months (October and November, 1995) and the analysis of the data took another month.

4. The descriptions of the oral test

The main curriculum, called Eigo Course , consists of six levels altogether. There are 18 units in each level. Upon completion of all units at each level, the students take a comprehensive test, which is the oral proficiency test. Each test is 7–10 minutes long with a native speaking interviewer and the interviewer takes the role of the judge as well. The purpose of the test is to measure the students' achievement at that level and also to diagnose the students' strengths and weaknesses at their present level. The tests mainly contain material which was actually taught in the class with

the exception of the intermediate, high intermediate and advanced levels, which include topics for free discussion (Refer to the Appendix A). The same criteria for evaluation are used for all levels and the judges give a holistic score as a percentage for the test result and the passing score is 70%. The five components for evaluation are word power, grammar, pronunciation, fluency and listening comprehension. The form doesn't include a weighting table with scales and it is used only when diagnosing the strengths and weaknesses of the students' performance. Rather, they have to give a more impressionistic score from a holistic perspective based on the students' achievements of the course content taught before the test. Also, the judges have to choose appropriate advice from the list compiled on the computer, based on the five components. An evaluation form is automatically computerized by inputting appropriate codes for advice (Refer to the Appendix B).

5. The validity and reliability of the test

Brown stated at the 1995 JALT conference that "language testing in Japan is very unscientific and in need of improvement." This test also might receive such a criticism because it was designed without thorough consideration as to the validity and reliability principles. First, the purpose of this test was not clearly decided. The test has two functions, that of an achievement test and that of a diagnostic test. The test covers mostly the curriculum which was taught at the level. However, whether the questions used in the test truly represent the achievement of the level is questionable because the selections of the questions are from a few selected units in each level. The final goal of the program is to improve the students' oral proficiency. However, the interpretation of the oral proficiency skill was not determined clearly by the test designers.

Various scholars have explored, reinterpreted or expanded the notion of communicative competence according to their own research interests (Murata, 1993). Among these, Hymes, Canale and Swain have played the most important roles. Hymes has extended to five, Chomsky's original two parameters of competence (1965), grammaticality and acceptability adding the three new domains of feasibility, appropriateness and actual occurrence (Hymes 1972, Munby 1978, Canale and Swain 1980, Savignon 1983, Widdowson 1983, 89). The theoretical definitions of communicative competence articulated by Canale and Swain were that effective communication relies on three types of competencies: grammatical competence, sociolinguistic competence and strategic competence. Canale reexamined them in light of other perspectives on language proficiency, and as a result he distiguished between sociolinguistic and discourse components of the communicative competence framework

(Canale 1983, Douglas and Chappelle, 1993). The test provides five components for evaluating oral proficiency, however none of the new concepts of communicative competence Hymes or Canale and Swain defined were taken into consideration by the test designers. The reliability of the test was estimated as follows. The same test was given to thirty students twice by two different judges (test–retest method) and the two sets of scores were statistically analyzed by the pearson product moment correlation and the Kuder–Richardson formula 20 (K–R 20). The correlation coefficient was .65, which is relatively low reliability and the standard error of measurement was 5.3. The unclear notions as to the purpose of the test and communicative competence are expected to increase the variations of the judges' marking.

6. The descriptions of the subjects and the students

Seven subjects were chosen for this research. Four judges had more than 2 year teaching experience and also the period of experience as a judge was also more than two years. The other three judges were relatively new and they had less than one year's experience as an EFL teacher and also as a judge. Table 1 shows the descriptions of the subjects. Protecting the participants involved guaranteeing that information obtained anonymously during the study from and or about the individuals will remain confidential. Researchers are not only ethically responsible to their subjects, but also to other constituencies (American Anthropological Association, 1970). Fifty students were chosen for this study from the levels of total beginners (15 students), beginners (15 students), Intermediate level (10 students), and advanced level (10 students).

7. The findings from observations and the analysis

A number of the tests were observed to collect the facts about the test, particularly to see how the tests were actually carried out and to see the similarities and differences in the methods and procedures the judges applied to the tests.

7.1. The significant influence of the first students' score on the other students' scores

There are three tests conducted per one hour. One interviewer is assigned to give three tests to three separate students. After careful direct observation it was found that the score of the first student had a great influence on the results of the other two students. It was observed that this tendency was commonly seen among all the judges who were observed. This phenomenon indicates that the first score from the first student is used as the criterion for marking during the hour of tests and the

Table 1 The descriptions of the subjects

	subject 1	subject 2	subject 3	subject 4	subject 5	subject 6	subject 7
nationality and sex	U.S.A. male	Canada male	U.S.A. female	Canada male	Canada female	U.S.A. male	U.S.A. female
length of teaching experience	3 years	2.5 years	2.5 years	2.5 years	1 year	6 months	8 months
length of experience as a judge	2.5 years	2 years	2 years	2 years	9 months	3 months	3 months
experience before teaching at this school	graduate school student	ESL teaching in Canada	teaching journalism in U.S.A.	elementary school teacher	under-graduate student	businessman	nurse

Table 2 The test scores

	The first score	The second score	The third score
Judge 1	82	83	65
Judge 2	76	74	68
Judge 3	75	80	no test given
Judge 4	77	75	63
Judge 5	65	68	65
Judge 6	70	70	no test given
Judge 7	90	92	86

other students' performances are compared with that of the first student. If the next student's performance doesn't differ too much from the first one, the scores closely resemble the first score in the manner of the scores seen in Table 2. In order to verify this phenomenon one of the questions in the interview given to these judges was the question regarding this point. All the judges answered that they might have been influenced by the first student subconsciously to set their own criterion. This shows that the first student performance is an important variable to affect the intrarater reliability and consequently affect interrater reliability in this test.

7.2. The different patterns of testing methods

It was found that each judge perceived, conceptualized and organized this oral interview differently even though they participated in the same training. The different patterns of testing methods were divided into two major groups shown in

Table 3 The variations in the testing procedure

	The first pattern	The second pattern
Starting and finishing time	No distinction between these two times. They made an effort to relax the students as much as they could. Some of them escorted the students to the testing room and began to have a relaxing conversation before the test.	They distinguished the testing time from before and after the test very clearly. They started the test by saying "I am going to start the test." and finished the test by saying. "The test is finished." They acted as a formal rigid tester.
Repetitions of delivering questions	They seemed to speak much slower than normal speed. They adjusted their speaking speed to the students' comprehension level, and many times ended up speaking very slowly.	They just repeated the same form of questions at the same speed only once. If the students didn't understand the questions, they moved on to the next question.
Speed of delivery	They slowed down when repeating or paraphrased the questions.	They repeated questions only once at the same speed.
Number of questions	If the students were weak and took more time to answer, most interviewers asked fewer questions to meet the time limit.	They always made the same number of questions no matter what the students' abilities were.
Ways of presenting questions	When moving on to the next question, some interviewers tried not to be abrupt and try to connect the questions in order to make natural conversation.	They asked the questions listed in the manual discretely.
Adjustment of the questions	When asking discrete questions using target structures, some interviewers tried to change questions slightly. Although they used the same structures, they changed questions so that the student could relate themselves to the questions.	They didn't make any changes of the model sentences using target structures and they just repeated the questions in the manual.

Table 3. The first pattern recognized in this study was the oral interview based on real-life, authentic naturally flowed conversation. The second pattern was the oral interview which adapted the mode of formal testing.

It is concluded that these distinct patterns are the main causes of variations in the judges' testing procedures. The inconsistency in the procedure is closely linked to the low reliability of the test. When obtaining the correlation coefficiency in the test -retest method, the majority of the students did better in the second test than the first test because the students were more relaxed in the second test than the first test (Practice effect). This indicates that the different atmosphere of testing which judges

create is a significant variable in changing the students' oral performance skill. Hughes (1989) suggests that testers who conduct oral tests should avoid constantly reminding candidates that they are being assessed. He also recommends that transitions between topics and between techniques should be made as natural as possible. His suggestions might contribute to improving the students' performance, and particularly if the interview is carried out as a natural conversation. In this case, the students finish the tests with a sense of accomplishment which can create a beneficial backwash.

7.3. Grading criteria

Grading procedure varied greatly from judge to judge. The following points were observed from direct observation and found to be significant:

7.3.1. Scoring procedure

The judges were not not provided with a weighting table, therefore each judge creates his or her own grading scheme. Six judges used holistic and impressionistic scoring and one did analytical scoring, based on the total of five components. The judges who have been conducting the test for more than two years (Judges 1,2,3,4) gave the final score immediately and confidently, however the judges who have been conducting the test for less than one year (Judges 5,6,7) took some time to compute the final score. The judges who have more experience in implementing the test seemed to have established their own formula to get the final score, however, the inexperienced judges seemed to be less confident in reaching the final score.

7.3.2. Different scoring criteria

Six of the judges gave scores such as 62, 67, and 72 and one judge used the numbers of 65, 75, and 95, occasionally rounding off to scores of 70, 80 and 90. The fractions appearing with the first group of judges have meaning only in that they usually appeared as extra points. The judge who didn't use such fractions felt it meaningless to give such numerical values.

7.3.3. Predictive validity

Three of the judges considered how well students would do at the next level (predictive validity) and others didn't consider this point at all.

7.3.4. Other factors affecting grading

Six judges took students' attitude into consideration. If the students acted

nervously, most of the judges felt sympathy and even though the students' performance was poor they added extra points. If they saw confident students who looked straight into the judges' eyes, they also added extra points to the total scores. One judge (Judge 3) tried not to be influenced by these factors.

8. The Analysis of the evidence from the interviews

In-depth Interviews were administered to these seven judges. The judges felt comfortable in answering the questions and many constructive opinions were contributed to improve the test quality. There were two purposes of the interview. One was to elicit information to provide different perspectives from the evidence found in the direct and taped observation. The other was to confirm the findings which were discovered in the previous observations. Analytic induction was used to analyze the transcribed interview data. In this approach the researcher returns repeatedly to transcripts to reread and reexamine the data, searching for salient or recurring themes (Johnson, 1992). The elicited information gave us an in-depth and emic descriptions of the mental mode of the judges. The evidence of the open-ended interview was categorized under the following points:

8.1. Testing time

It is unlikely that much reliable information can be obtained in less than about 15 minutes, while 30 minutes can probably provide all the information necessary for most purposes (Hughes, 1989). However, almost all the judges who were interviewed said that the test time (7 minutes) was basically long enough except for one judge who preferred a longer test. She stated as follows.

1. J2: I'm trying to have natural conversation with the students during the test. I have to spend the first couple of minutes to relax them. Lower level students may need a longer time than the higher level students. Also, for the students who are extremely nervous, a longer time would help them to relax them.

8.2. *Understanding the purpose of the test*

Even though they took part in the same training, the judges seemed to have gotten different ideas about the purpose of the test and what is to be measured by the test. The different answers for (1) the purpose of the test and (2) what is to be measured by the test are significant to the variances in the evaluation criteria. The four judges who had been working for more than two years gave clear definitions of the two questions, however the three judges who had relatively short experience in

administering the test seemed to be less confident in answering these two questions.

2. J1: The most important purpose of the test is to see if the students can succeed at their present level, so my interpretation on the test is that it is a diagnostic test., but first I judge if the student will be able to handle the next level or not, then if he or she is good enough, I give more than 70% as my final score of the test depending on their performance. After that, I pay attention to the five components and evaluate the students' strong and weak points based on these points.

3. J2: The purpose of the test is to find out how much material covered in class has been mastered by the student, so I think it's an achievement test. I tried to measure their conversation ability which involves grammar, syntax, verb usage, particularly tenses, and manipulation of various structural patterns.

4. J3: I try to determine whether the students had mastered the grammar at their present level and whether they possess the comprehension skill to handle the next level. The most important measuring criteria are listening comprehension , grammar for the level and fluency (appropriateness, speed, and rhythm).

5. J4: The purpose of giving the test is to check the student's comprehension of the objectives of the lessons and their ability to use them. I think the five components provided are all important, but it's difficult to compute each factor accurately.

6. J 6: I had a hard time figuring out what this test was looking for. Now I know the test is trying to determine whether the students have achieved the minimum level of the book. I think the five components are useless in determining the final decision of whether they pass or fail, but they are useful in giving specific comments to the students.

8.3. Possible threats to the reliability

Several questions were asked to find out whether certain psychological factors that the judges have might affect them actually work as threats in assessment. The followings are the factors and the results of the judges' answers to the questions.

8.3.1. The familiarity with the students' performance

If the judges know the students and their performance in class relatively well, it affected their performance. Although the judges are not supposed to take the students classroom performance into consideration in assessment, most of the judges said that it was possible to be influenced, especially when the students did perform

well in class but didn't on the test.

7. J5: If I know the student very well, it's easy to assess the student' performance. Sometimes I already make up my mind before I give a test. I know I shouldn't do that.

8.3.2. The students' attitude

Students' attitude was one of the biggest factors which influenced the scores. The various factors appeared in their answers.

8. J1: The student's attitude itself doesn't affect my assessment. However, it affects the student's performance, particularly fluency.

9. J4: Nervousness contributes to some extra points in my case, but extreme nervousness affects adversely.

10. J6: Nervousness contributes to extra points. If the student is a little below or at the border line, I might pass the student.

11. J 7: If a student is very confident, I will give extra points. I think the attitude is a very important factor to have successful communication.

9. Conclusion and Applications

In this study various factors which affected the judges' evaluation were found. Although it is usually impossible to achieve a perfectly reliable oral test, test constructors must make their tests as reliable as possible. They can do this by reducing the causes of unsystematic variations to a minimum. The findings in this study reveals those unsystematic variations are the cause of low reliability and clarifying the marking scheme of the test will make the test more valid and reliable. The judges cannot foresee all of the responses that candidates might come up with as answers to their questions correctly, therefore the training for judges needs to be implemented thoroughly and the marking scheme should be explained in as much detail as possible. Whereas pencil –and –paper test types depend largely on statistical procedures for determining their validity and reliability, interview–based tests depend heavily on the quality of training of the interviewers and raters (Douglas and Chappelle,1993). This study indicates that only well trained interviewers should administer the tests. A monitoring system should be established to ensure that the quality of interviewing and rating is maintained. Also the test administrators should consider any negative effects on the quality of interviewing and rating. If we apply

analytical criteria to the spoken product of test tasks the issue still remains of what the profile of achievement of a successful candidate is. In other words, we have to be explicit about the level of performance expected in each of the specified criteria. In addition, there is a question mark hanging over analytic schemes. Carrol (1961, 1972) recommends tests in which there is less attention paid to specific structure points or lexicon than to the total communicative effect of an utterance. However, one potential advantage of the analytical approach is that it can help provide insight into a candidate's weaknesses and strengths which may be helpful diagnostically, and also make a formative contribution in course design (Weir,1990). Also FSI applies the weighting table developed through experimentation that has the heaviest emphasis on grammar, secondary emphasis on vocabulary, and the least emphasis on accent. These weightings permit the numerical total score to correspond to the levels of proficiency (Refer to the appendix C). Such an experimentally verified usage of analytic schemes greatly contributes to the interrater reliability.

A comparison of test specification and test content is the basis for judgments as to content validity. A set of specifications for the test, is the information such as content, format and timing, criteria levels of performance, and scoring procedures. The specifications should accurately represent the characteristics, usefulness and limitations of the test, and describe the population for which the test is appropriate (Alderson, Clapham, Wall,1995). Also the definition of oral proficiency should be clarified in the specifications. The underlying problem in considering proficiency scales and tests is defining what is meant by proficiency itself. Farhady (1982) states that language proficiency is one of the most poorly defined concepts in the field of language testing. Nevertheless, in spite of differing theoretical views as to its definition, a general issue on which many scholars seem to agree is that the focus of proficiency is on the students' ability to use language. A number of definitions have focused solely on the use of the language. Clark (1975), for example, has defined proficiency as the ability to receive or transmit information in the test language for some pragmatically useful purpose within a real-life setting. With this view the interview should be a naturally flowing conversation between the interviewer and the student.

Finally, to maximize reliability, the interview should be tightly structured to control what the interviewer can do, and a process of monitoring the interview quality and rating accuracy should also be built into administrative proceedings.

References:

Alderson, C. J., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation.*

Cambridge: Cambridge University Press, 20–24.

American Anthropological Association. (1970). Principles and professional responsibility. Newsletter, 14–16.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative competence. *TESOL Quarterly,* 16, 449–465.

Brown, J. D. (1995). The Testing in Japan. The 21st JALT International Conference. Nagoya: Japan.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication.* London: Longman, 2–27.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics,* Volume 1, No. 1, 1–47.

Carrol, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. Reprinted in H. Allen and R. Campbell (eds.) 1972, *Teaching English as a second language: a book of readings.* New York: McGraw–Hill.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Mass: The M.I.T. Press.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In Jones, R.L., & Spolsky, B. (Eds.), *Testing Language Proficiency.* Arlington: VA., Center for Applied Linguistics.

Douglas, D. & Chapelle, C. (1993). *A new decade of language testing research.* Virginia: Teachers of English to Speakers of Other Languages, Inc., 222–234.

Falk, B. (1984). Can grammatical correctness and communication be tested simultaneously?, *Practice and problems in language testing.* Oxford: Oxford University Press.

Farhady, H. (1982). Measures of language proficiency from the learners's perspective. *TESOL Quarterly,* 16, 43–61.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Hymes, D. (1972). On communicative competence. In Pride, J.B. and Holmes, J. (eds), *Sociolinguistics: Selected Readings.* Harmondsworth: Penguin, 269–293.

Jafarpur. A. (1988). Non–native raters determining the oral proficiency of EFL learners. Department of Foreign Languages, Shiraz University, Iran. System, Vol. 16, No. 1, 61–67.

Johnson, M. D. (1992). *Approaches to research in second language learning.* London: Longman, 86, 101.

Lado, R. (1961). *Language testing.* London: Longman.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

Murata, K. (1993). Communicative competence and capacity: What's the difference? : A Critical review. *JACET, Kiyo,* Vol. 24.

Savingnon, S. (1983). *Communicative competence: Theory and classroom practice.* Mass: Addison Wesley.

Weir, C.J. (1990). *Communicative language testing.* Englewood. Cliffs. NJ: Prentice–Hall Regent.

Widdowson, H.G. (1983). *Learning purpose and language use.* Oxford: Oxford University Press.

Appendix A

Oral Interview Test: Question Samples

(Level 1–Total Beginners, Level 2–Beginners, Level 4–Intermediate level, Level 6–Advanced level)

Level 1 Section 1: Personal Questions

1. What is your name?
2. What do you do?
3. How many members are there in your family?
4. What do you study at your university?
5. What is your hobby?

Section 2: Role play

Introduce yourself to someone you don't know.

Level 2 Section 1: Personal Questions

1. Tell me about your family.
2. What club do you belong to?
3. Have you ever been to a foreign country?

Section 2: Role Play

Explain how to get to Osaka station to a foreign tourist in the street.

Level 4 Section 1: Personal Questions

1. Introduce yourself.
2. What would you like to be in the future?
3. Why do you want to master English?

Section 2 : Introduce some Japanese food and customs to a foreign guest in a Japanese restaurant.

Level 6 Section 1: Personal Questions

1. Introduce yourself.
2. Are you a self-assertive person?
3. Explain what you do at your company.

Section 2: Free Conversation

1. What contribution should Japan make to world peace?
2. What do you think of the Japanese educational system?
3. What do you think of the status of women in Japan?

Appendix B

Student's Name _____ Student's Level _____

Proficiency Level % Pass Fail

Comments

Word Power _____

Grammar _____

Pronunciation _____

Fluency _____

Listening _____

General Comments

Interviewed by

Date

_____ _____

Appendix C

FSI Weighting Table

Proficiency Description	1	2	3	4	5	6	
Accent	0	1	2	2	3	4	_____
Grammar	6	12	18	24	30	36	_____
Vocabulary	4	8	12	16	20	24	_____
Fluency	2	4	6	8	10	12	_____
Comprehension	4	8	12	15	19	23	_____

The total score is then interpreted with the Conversion Table that follows:

ESL Conversion Table

Total Score	Level	Total Score	Level	Total Score	Level
16–25	0+	45–52	2	73–82	3+
26–32	1	53–62	2+	83–92	4
33–42	1+	63–72	3	93–99	4+