ASIAN EFL JOURNAL
The EFL Professional's Written Forum

ISSN: 1738-1460

Volume 11. Issue 1
Article 8

Title

Foreign Language Speaking Assessment: Chinese Taiwanese College English Teachers' Scoring Performance in the Holistic and Analytic Rating Methods

Author
Ying-Ying Chuang

Bio Data:
Ying-Ying Chuang is an assistant professor in the Department of Applied Foreign Languages at Cheng Shiu University, China Taiwan. She holds Ed.D. in Bilingual Education from Texas A&M University. She is currently teaching the courses of Second Language acquisition, English Reading, and ESL Teaching Methodologies. Her research interests include second/foreign language assessment and language learning strategies.

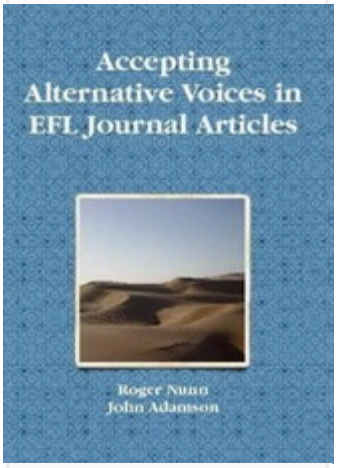
Abstract
The purpose of this study was to investigate college English teachers' scoring performance of the holistic and analytic rating methods, their views and concerns with the components of oral skills, and whether teachers' background variables influenced their scoring performance. Compared with the individual teacher's rating scores, the results indicated that no statistically significant differences were found between the scores in two rating methods. The majority of the teachers ranked "comprehensibility" as their most concern of oral performance, while "vocabulary/word choice" happened to be the least of their concern. Regarding the relationship between rated scores and *rater effect*, the findings indicated that the statistically significant differences were found in the factors of the teachers' age and academic major; however, no statistically significant differences were found in the factors of teaching experience and rating training. Some pedagogical implications of the study are included for further inquiry.

Keywords: oral proficiency; speaking assessment; rating scale; English as foreign language.

Introduction

Background of the Study

Speaking seems intuitively the most important of all the four language



skills (listening, speaking, reading, and writing) since people who know a language are referred to as a 'speaker' of that language, as if speaking included all other skills of knowing that language (Ur, 1996). In other words, a learner's end product of language learning is to be capable of speaking the target language fluently. However, speaking skill is a crucial part of the language learning process, and it is also the one skill, which has often been neglected in EFL classroom. In addition, in English proficiency testing, oral performance appears to be one of the most difficult skills to assess since there are many external and internal factors that influence a rater's impression toward how well someone can speak a foreign language. In other words, the reliability of scoring has always been doubted as the oral proficiency test inevitably involves raters' personal/ subjective views instead of their objective points of view.

Based on the fact that many learners who had been taught the English language by nonnative speakers in the countries that consider English as a foreign language, Kim (2005) claimed that using the rating criteria based on native speakers' standards to measure learners' oral proficiency was not appropriate for the actual use of English in an international context. Therefore, it is important for educators, test designers, and researchers to reconsider the purposes of language speaking tests, and the standards of assessing learners' speaking skills, since it cannot be denied that the natural function of speaking is for a meaningful message delivery rather than the use of language form.

Statement of the Problem

Recently, many English educators in Taiwan have promoted curriculum reforms in order to meet the principles of Communicative Language Teaching. In light of this, a great deal of attention has been focused on revising teaching materials and curriculum, which were meant to improve teaching facilities for the attainment of communicative goals. However, the idea that teachers should improve evaluations by promoting the communicative approach has been neglected. Nevertheless, even students can receive good grades in English courses; however, it does not mean that their oral proficiency has achieved a certain level of competency. In addition, most assessments taking place in the Taiwanese English classrooms were conducted by pencil-paper tests without considering the importance of oral production in language learning (Cheng, 2006). The reasons that teachers avoided doing oral tests include the amount of time it took, the large size of student population, and students' negative reactions toward oral testing (Liu, 2006; Teng, 2005).

From the learner's perspective, speaking test is the most complex and difficult task among the language skills since their preparations should include knowledge about the language and the skills involved in using it (Bygate, 1987). Wang (2003) conducted a survey of Taiwanese college students in freshman English classes and she noticed that within the four language skills, speaking ability was the one that the students thought they should improve the most (83.7%). This meant that many students thought their oral skill was deficient. Also, some studies indicated that Asian students indeed had comparatively high anxiety in English learning (Na, 2007; Tsai, 2003) since most of them lacked speaking practice in the target language

both inside and outside of the classroom. This limited real-life practice and experience appears to have eroded their confidence and weakened their willingness to speak. Moreover, they experienced a sense of panic when pressured into doing an English oral test.

Purposes of the Study

Speaking assessments have become one of the most central issues in language testing. Unfortunately, few studies had been completed with the focus on foreign language speaking skills in Taiwan (Chen, 2001; Li, 2003; Lin, 1996; Pan, 2002; Wang, 2003), and the available research in the field of EFL speaking assessment is inadequate. Therefore, the present study aimed to investigate: how the rated scores differed between the holistic and analytic rating methods when the English teachers assessed the Taiwanese college student speech samples, if the teachers' characteristics affected their rating performance, and their understanding along with major concerns with the components of oral production. The researcher believes that it is necessary for English educators in Taiwan to rethink the questions of how and what to *assess* in speaking in order to help learners improve their oral skills, and particularly to be aware of the impact from *washback* (Bailey, 2005), to see the effects of testing on teaching and learning.

Research Questions

This study addressed the following research questions:
1. How did the rated scores differ between the holistic and analytic rating scales when English teachers assessed the same five Taiwanese college students' speech samples?
2. Was there any difference between English teachers' rankings of the five components—grammatical accuracy, vocabulary/word choice, pronunciation/accent, flow of speech/fluency, and comprehensibility—in the analytic rating scale when they assessed EFL student oral language proficiency?
3. How did the teachers' background characteristics—age, academic major, and teaching experience—influence their rating scores of the speech samples?

Review of literature
New Direction of Language Testing
Traditionally, language testing has taken the form of testing knowledge about the language: grammar and vocabulary. However, there is much more to using a language than just knowledge about it. Hymes (1974) argued that a language learner should not only have the ability to form correct sentences, but also to use them at appropriate times. The main purpose of communicative language tests is to assess the test taker's ability to use the language in real-life situations. In testing speaking skills, the focus should center on producing the appropriate and meaningful messages rather than grammatical accuracy (Kitao & Kitao, 1996). For instance, for those EFL learners who learn the target language for specific purpose situations, the tests should reflect what they actually need and what is useful to apply in those specific communication situations, such as occupational or professional areas. While some learners do not have a specific purpose—such as those

students who learn English as a required academic subject—the language tests for them can be directly focused on general social situations where they might have the chance to use English (Kitao & Kitao, 1996).

With the ever-increasing popularity of Communicative Language Teaching (CLT), language testing has come to view communicative purpose as its central concern. For example, the TOEFL test (Test of English as a Foreign Language) has now undergone a major makeover—the biggest change was to add a new speaking component, beginning in year 2006—aimed at better evaluating how well applicants can orally communicate in the target language (English). The English Testing Service (ETS) explained that the old test version failed to identify those students who mastered only 'textbook' English, and the educators of ETS hope that the change will improve English language teaching and learning worldwide, particularly of those students from Asia, where schools generally emphasize vocabulary, grammar, and reading skills over speaking skills in second/foreign language teaching and learning (Cheng, 2006; Liu, 2006).

Communicative language tests are those which make an effort to test language in a way that reflects the way that language is used in real communication; they focus on language meaning and function rather than language form. If students are encouraged to learn the target language through more communicative ways, it would make a positive effect on their language learning.

Variability in the Assessment Process: The Role of Rater

The issue of test features may influence a certain and significant degree of impact towards test results, particularly the rater effects, such as the influencing factors such as age, native language, gender, attitudes, professional background and experience, and so forth. Rater effects must be considered if the rating score is to be accurate of a learner's language performance, and to reflect a learner's real language ability in a test. According to Brown (2004), oral proficiency tests usually employ human raters to judge and score a test-taker's performance. The role of the raters is extremely important in the process of oral language assessments. Not only does their professional judgment impact decision-making in scoring, but their reliability also influences the meaning and quality of the scores.

Some previous studies have examined the relationship between raters and test scores in ESL oral proficiency assessments (Bachman et al., 1995; Brown, 1995; Lumley & McNamara, 1995; Caban, 2003; Kim, 2005). Most of their findings have proven that a test rater's background did affect their rating behaviors in certain aspects and those differences in background brought different scores from the raters.

Variability in the Assessment Process: The Role of Rating Scales

The test scores of oral language proficiency reflect how well the learner can speak the language being tested on a rating scale. Davies et al. (1999) defined the rating scales:

A scale for the description of language proficiency consisting of a

series of constructed level against which a language learners' performance is judged….the levels or bands are commonly characterized in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion) ….raters or judges are normally trained in the use of proficiency scales so as to ensure the measure's reliability. (pp. 153-154)

The rating scale for speaking was made up of an ascending series of levels, and each level should provide a statement as a scale descriptor to describe what each level or score meant. North (2000) described the challenge of developing a rating scale as trying to describe the complexity of a language ability in a small number of words. There were different types of rating scales that could be employed to score learners' speech samples; one of the traditional distinctions was between the holistic and analytic rating scales (Fulcher, 2003). A holistic rating captured an overall impression of the speaker's performance: a primary trait score assessed the speaker's abilities to achieve a specific communication purpose. In a holistic scoring, the rater reacted to the speaker's oral production as a whole: one score was awarded for his or her speech performance. Normally, this marked score was on a scale of 1 to 5, or even 1 to 10. Often each level on the scale was accompanied by a verbal description of the performance required to achieve that score (score criteria).

On the other hand, an analytic rating assessed and captured the speaker's performance on a variety of categories, such as delivery, organization, content, and language. The analytic categories, which the test developer included in his or her rating system amounted to his or her theory or hypothesis of what speaking was about. Some people agreed that the holistic rating was desirable for the evaluation of the general communicative effectiveness of the test-taker, however, "raters can be confused when evaluating many things simultaneously" (Kim, 2005, p. 52). The analytic rating tended to identify sub-skills such as grammar, vocabulary, pronunciation, and fluency. Generally speaking, the holistic scales were more practical for decision-making since the raters only marked one score: the flexibly allowed many different combinations of strengths and weakness within a level. From a rater's perspective, "holistic rating scales make [scoring easier and quicker] because there is less to read and remember than in a complex grid with many criteria" (Luoma, 2004). However, the advantages of the analytic rating were due to the detailed guidance that was offered to the raters, and the rich information as criteria was provided on specific strengths and weaknesses of the test-taker's performance. Therefore, Fulcher (2003) pointed out that to select the most appropriate type of rating scale for a particular speaking test, either holistic or analytic, the purposes of the test should be an element of the decision. Bachman and Savignon (1986) also suggested that a holistic rating, along with an analytic rating, should be assigned to provide a precise profile of the examinee's speaking ability.

In conclusion, identifying an appropriate rating scale depends upon the purposes of the assessment, and the availability of existing instruments. Rating systems may describe varying degrees of competence along a scale or may indicate the presence or absence of a characteristic. As Weigle (2002) mentioned, the choice of testing

procedures should involve finding the best possible combination of the qualities (reliability, validity, and so forth) and deciding which qualities were most relevant in a given situation. Therefore, a major aspect of any rating system is rater objectivity. The reliability of raters should be established during their training and checked during administration or scoring of the assessment.

Methodology and Procedure

Instrumentation

This study aimed to compare the two types of rating methods employed by teachers to evaluate EFL students' oral performance, and to investigate the analytic components of oral production. To gather the information from the raters, this study employed a rater survey and two types of rating instruments (holistic and analytic) for quantitative method analysis, and a rater interview as a qualitative method to support quantitative data.

The Rater Survey included the items which related to teachers' personal background information, such as age, native language, academic major, teaching experience, experience of rating oral proficiency and using rating scales, and whether they had been trained for rating oral proficiency. These demographic items provided the raters' similarities and differences in order to determine if these characteristics influenced their rating behaviors.

All selected raters needed to assess the five Taiwanese college student speech samples using the holistic and analytic rating scales which were originally designed by Kim (2005) but revised by the researcher. The holistic rating method, according to Kim, was designed to "ask raters to provide a score for the overall impression of the speaker's English language oral proficiency without any specified rating criteria" (p. 60). Another rating method, the analytic rating scale, asked all raters to score five rating competences: grammatical accuracy, vocabulary/word choice, pronunciation/accent, rate of speech/fluency, and comprehensibility. The raters scored the speech samples from a level of 1 to 7 to present their opinions from 'low proficiency' to 'high proficiency' in each component.

In this study, individual interviews were also conducted. The researcher further contacted the raters who were willing to answer the specific questions from the survey by using the telephone and face-to face interview in order to clarify their responses, expose their beliefs, and provide detailed opinions.

The Subjects of the Study

The researcher chose the teachers who have taught English courses at the universities in southern region of Taiwan, to serve as raters in this study. They were asked to mark scores to the speech samples by using the two rating scales, and then completed a teacher survey.

Data Collection

The researcher sent each selected rater an e-mail with the files, including 1) the letter to the subject with the instruction for rating speech samples, 2) five student speech samples, 3) a rating booklet which included the analytic rating scale descriptor, the holistic and

analytic rating scoring forms, and 4) the Rater Survey. The university-level English teachers in southern Taiwan were selected and submitted materials. Further, in order to obtain insights from the raters, individual interviews with the raters were also arranged.

Findings

Demographic Data

A total of 80 copies of the survey and rating instruments were sent, of which 62 copies were returned. Therefore, the overall response rate was 77.5%. Regarding respondents' gender, fourteen raters (22.6%) were male, and forty-eight (77.4%) were female. Nearly one-fifth of raters were at the age range of 21-30 (19.4%); over half of the raters were at the age range of 31-40 (56.5%); 14.5% of the raters were at the age range of 41-50; and 9.6%of the raters identified their age range as over 50. The raters' academic majors were varied. The researcher categorized them into three fields: linguistics/English literature, TESOL/ESL education, and "others." There were nine raters who had an education background in linguistics/English literature (14.5%), nearly two-thirds of raters with TESOL/ESL education background (61.3%), and nearly one-fifth of raters belonged to the group of "others" (24.2%).

Findings and Discussion of Research Question One

Research Question One: ***How did the rated scores differ between the holistic and analytic rating scales when English teachers assessed the same five Taiwanese college student speech samples?***

During the data collection process, each rater was required to rate the same five speech samples twice: first time to assess the speech samples by using holistic rating scales, and second time by using analytic scales. Descriptive statistics of the five speech samples describes the characteristics of a score distribution rated by the raters. The findings are presented in Table 1.

Table 1
Descriptive Statistics of Overall Holistic Ratings from Student Speech Samples (N=62)

Speech Sample	N	M	Min	Max	Range	SD
1	62	4.34	2	7	5	.96
2	62	4.77	3	7	4	.93
3	62	5.24	2	7	5	1.13
4	62	4.37	2	7	5	1.09
5	62	3.37	1	6	5	1.06

The mean range of the holistic rating scores rated by the raters was from 3.34 to 5.24. Speech sample 3 received the highest mean ratings among the five speech samples, and speech sample 5 received the lowest mean ratings. In addition, the mean scores of speech sample 1 and 4 were very close, the difference was only 0.03.

Table 2
Descriptive Statistics of Overall Analytic Ratings from Student

Speech Samples

Speech Sample	N	*M*	Min	Max	Range	*SD*
1	62	4.55	2.40	6.80	4.40	.92
2	62	4.55	3.00	6.20	3.20	.79
3	62	5.26	3.00	7.00	4.00	.93
4	62	4.47	2.40	7.00	4.60	.95
5	62	3.46	1.00	5.80	4.80	1.02

Table 2 describes descriptive statistics of the raters' scores for analytic ratings of each speech sample. The lowest range obtained is 1 and the highest is 7, indicating that raters used the whole ranges of the rating scales. The total raters' mean range of analytic ratings is from 3.46 to 5.26. These were slightly higher than their holistic ratings. The same results were found as in holistic ratings. Speech sample 3 received the highest mean score among the raters' analytic ratings, and speech sample 5 received the lowest mean score of the ratings.

Table 3
Descriptive Statistics of Overall Holistic and Analytic Scores (N = 62)

Rating Scale	*M*	*N*	*SD*	Std. Error Mean
Holistic Scores	4.44	62	.71	.09
Analytic Scores	4.47	62	.70	.09

The overall mean scores of holistic and analytic ratings were compared to examine the score differences between the two ratings results. The data in Table 3 indicated that the mean score of the holistic ratings ($M = 4.44$, $SD = 0.71$) from the sixty-two raters was slightly lower than their mean score of the analytic ratings ($M = 4.47$, $SD = 0.70$).

In order to evaluate whether there were statistically significant differences between the two rating methods, a paired-samples *t*-test was conducted.

Table 4
Paired-Samples t-Test of Overall Holistic and Analytic Scores (N = 62)

Holistic Scores	Paired Differences				
-	*M*	*SD*	*t*	*df*	*Sig.* (2-tailed)
Analytic Scores	.03	.41	.54	61	.59

The findings in Table 4 indicate that the mean of the holistic scores ($M = 4.44$) was close to the mean of the analytic scores ($M = 4.47$), $t(61) = 0.54$, $p = 0.59 > 0.05$. Therefore, the findings revealed that there was no statistically significant difference found between the rated scores of the two rating scales.

Findings and Discussion of Research Question Two

Research Question Two: ***Was there any difference between the***

English teachers' rankings of the five components—grammatical accuracy, vocabulary/word choice, pronunciation/accent, flow of speech/fluency, and comprehensibility—in the analytic rating scale when they assessed EFL students' oral language proficiency?

Item 14 of the Rater Survey focused on the rater's opinions of the components of the learners' oral proficiency competence based on their opinions. The raters were asked to rank each of the five components—grammatical accuracy, vocabulary/word choice, pronunciation/accent, flow of speech/fluency, and comprehensibility—followed by the number from 5 (the most important) to 1 (the least important). The raters' rankings were compared by frequency.

Table 5
Frequency of the Raters' Response for Ranking the Most Important of the Five Components

Analytic Component	Frequency (n)	Percentage (%)
Grammatical Accuracy	9	14.5
Vocabulary/Word Choice	5	8.1
Pronunciation/Accent	11	17.7
Flow of the Speech/Fluency	5	8.1
Comprehensibility	32	51.6
Total (N)	62	100.0

Table 5 depicts the frequency of the raters' responses for their ranking the most important of the five components in the analytic rating scales. The majority of the raters ranked "comprehensibility" as their first concern when they assessed student oral language proficiency (51.6%), followed by the "pronunciation/accent" (17.7%). However, there were only five raters who answered that they considered "flow of the speech/fluency" or "vocabulary/word choice" as the most important component of the speaking abilities (8.1%).

Table 6
Frequency of the Raters' Response for Ranking the Least Important of the Five Components

Analytic Component	Frequency (n)	Percentage (%)
Grammatical Accuracy	11	17.7
Vocabulary/Word Choice	23	37.1
Pronunciation/Accent	12	19.4
Flow of the Speech/Fluency	4	6.5
Comprehensibility	12	19.4
Total (N)	62	100.0

On the other hand, the raters' responses, which showed their opinions on ranking the least important of the five components, are described in Table 6. More than one-third of the raters (37.1%) considered that "vocabulary/word choice" was the least important of the five components. This meant that when those teachers rated students' oral proficiency, their scores might not reflect the students' competence in

vocabulary used. In other words, teachers might pay more attention to students' oral performance in other components, such as if they are able to express their meanings clearly, to apply correct grammatical rules, and to pronounce accurate sounds. Table 6 also indicated that only 6.5% of the raters ranked the "flow of the speech/fluency" as their last choice when they assessed students' oral performance.

Table 7
Frequency of Response for Ranking the Most Concern of the Five Components Based on Rater Variables

Rater Variable	G	V	P	F	C	Total
Native Language						
English	2	1	2	-	4	9 (14.5%)
Mandarin Chinese	7	6	10	5	25	53 (85.5%)
Academic Major						
Ling. /Literature	1	2	2	1	3	9 (14.5%)
TESOL	6	3	7	4	18	38 (61.3%)
Others	2	2	3	-	8	15 (24.2%)
Years of Teaching Experience						
Less than 2 years	2	-	1	1	7	11 (17.7%)
3 to 6 years	5	4	6	4	16	35 (56.5%)
Over 7 years	-	3	5	-	6	16 (25.8%)
Rating Experience						
Less than 2 years	7	3	7	3	23	43 (69.3%)
3 to 6 years	-	2	3	1	6	12 (19.4%)
Over7 years	2	-	1	1	3	7 (11.3%)
Teaching Certificate						
Yes	6	5	7	4	17	39 (62.9%)
No	3	2	5	1	12	23 (37.1%)
Rating Training						
Yes	2	5	-	2	9	21 (33.9%)
No	7	2	9	3	20	41 (66.1%)

Note. G = Grammatical Accuracy; V = Vocabulary/Word Choice; P = Pronunciation;
F = Flow of the Speech/Fluency; C = Comprehensibility.

In order to find out if raters' background variables influenced their rankings, the researcher divided the raters into groups according to their native language, academic major, teaching and rating experience, and having or not having English teaching certificates and rating training. Table 7 revealed that in all groups with different variables, "comprehensibility" received the highest ranking. Regarding the raters' native language, the majority of the teachers believed that "comprehensibility" was their first concern when they checked students' oral performance.

However, the raters who were with different academic backgrounds, with or without rating experience, having or not having English teaching certificates or rating training, showed no difference with their selections in general, including the less- and well-experienced teachers. This meant that when teachers test students' oral proficiency, they would focus more on checking whether their speeches can be understandable and intelligible. Moreover,

"pronunciation/accent" received the second highest ranking. However, none of the raters who had been trained in rating speaking assessment selected "pronunciation/accent" as their priority.

Compared with the different background variables of the subjects, the raters' responses, which showed their opinions on ranking the least important of the five analytic components, are described in Table 8.

Table 8
Frequency of Response for Ranking the Least Concern of the Five Components Based on Rater Variables

Rater Variable	G	V	P	F	C	Total
Native Language						
English	2	4	1	1	1	9 (14.5%)
Mandarin Chinese	9	17	11	3	13	53 (85.5%)
Academic Major						
Ling. /Literature	3	3	2	-	1	9 (14.5%)
TESOL	6	13	8	2	9	38 (61.3%)
Others	2	5	2	2	4	15 (24.2%)
Years of Teaching Experience						
Less than 2 years	1	4	3	2	1	11 (17.7%)
3 to 6 years	5	11	8	2	9	35 (56.5%)
Over 7 years	5	6	1	-	4	16 (25.8%)
Rating Experience						
Less than 2 years	6	16	10	4	7	43 (69.3%)
3 to 6 years	3	3	2	-	4	12 (19.4%)
Over7 years	2	4	-	-	1	7 (11.3%)
Teaching Certificate						
Yes	7	17	5	3	7	39 (62.9%)
No	4	6	7	1	5	23 (37.1%)
Rating Training						
Yes	4	7	3	1	6	21 (33.9%)
No	7	14	9	3	8	41 (66.1%)

Note. G = Grammatical Accuracy; V = Vocabulary/Word Choice; P = Pronunciation;
F = Flow of the Speech/Fluency; C = Comprehensibility.

The data showed that the majority of the raters in each variable group preferred "vocabulary/word choice" to the other four analytic components, except for the group of raters with 3 to 6 years rating experience of speaking assessment. Particularly, half of the raters in that group who had taught English for more than 7 years did not focus on "vocabulary/word choice" during the process of rating their students' oral production. Table 8 also indicated that only few raters selected "flow of speech/fluency" as the least important among the five components. It appeared that even if most raters did not prioritize "flow of speech/fluency," it did not mean that the student's performance in fluency would be totally neglected by those raters in speaking assessment.

Regarding how the teachers ranked their components in sequence as well as why they made such decisions, the information gathered from interviews can provide more insights and details.

Findings and Discussion of Research Question Three

Research Question Three: ***How did the teachers' background characteristics—age, academic major, and teaching experience—influence their rating scores of speech samples?***
From Research Question One, the findings indicated that the raters' rating scores had no statistically significant difference between the two rating methods. It meant that the means of the rated scores in those two rating scales were very similar. Therefore, only the holistic rating scores were used to evaluate whether the rated scores were impacted by the raters' age, academic major, and English teaching experience to Taiwanese college students. The responders were divided into four age groups: 21 to 30, 31 to 40, 41 to 50, and over 50. The descriptive statistics are illustrated in Table 9.

Table 9
Descriptive Statistics of Holistic Scores Based on Raters' Age (N = 62)

Age	*N*	*M*	*SD*
21-30	12	3.85	.50
31-40	35	4.59	.66
41-50	9	4.53	.64
Over 50	6	4.60	.83
Total	62	4.44	.70

The raters' lowest mean score was at the age group of 21 to 30 (M = 3.85, SD = 0.50), while the highest was at the age group of over 50 (M = 4.60, SD = 0.83). The raters in the age group of 21 to 30 were the strictest since they were the only group whose mean score was lower than the overall mean score of the raters (M = 4.44, SD = 0.70) in the holistic ratings. The raters of the age group over 50 were the most lenient out of the four groups.

A one-way analysis of variance (ANOVA) was conducted to evaluate the relationship between rated scores and the raters' age levels. The independent variable, the rater factor, which included four age groups. The dependent variable was the rated holistic rating scores.

Table 10
ANOVA Results for the Effect of Raters' Age on Holistic Scores (N = 62)

Source	*df*	*SS*	*MS*	*F*	*Sig.*
Between Groups	3	05.24	1.75	4.19	.009*
Within Groups	58	24.21	0.42	-	-
Total	62	29.45		-	-

Note. * $p < 0.05$

As the findings showed in Table 10, there was a strong relationship between the age factor and the overall holistic scores rated by the raters since there were statistically significant differences found between the age groups in holistic scores at $p < 0.05$ level (p = 0.009).

However, using ANOVA alone could not determine which age levels of the raters differ from each other in the holistic rating scores they rated. A post hoc comparison test would be employed to decide precisely which age group means were significantly different from other group means.

Table 11
Post Hoc Multiple Comparisons for Rated Holistic Score Based on Raters' Age (N = 62)

(I) Age	(J) Age	Mean Difference (I-J)	Std. Error	*Sig.*
21-30	31-40	-.74	.22	.00*
	41-50	-.68	.29	.02*
	Over 50	-.75	.32	.02*
31-40	21-30	.74	.22	.00*
	41-50	.06	.24	.80
	Over 50	-.01	.29	.98
41-50	21-30	.68	.29	.02*
	31-40	-.06	.24	.80
	Over 50	-.07	.34	.85
Over 50	21-30	.75	.32	.02*
	31-40	.01	.29	.98
	41-50	.07	.34	.85

Note. *$p < 0.05$

Tested by post hoc multiple comparisons of group means using the LSD method, the findings in Table 11 reveal that there are statistically significant differences between the raters of the age group 21 to 30 and the age group of 31 to 40 ($p = 0.00 < 0.05$), 41 to 50 ($p = 0.02 < 0.05$), and the raters in the age group of over 50 ($p = 0.02 < 0.05$). It meant that the overall holistic scores rated by the raters at the age group of 21 to 30 were significantly stricter than the other age groups.

The rater's academic major factor also was tested to see if any relationship existed with the rated holistic rating scores. The responders were divided into three groups: major related to linguistics or English literature, major related to TESOL or ESL education, and others.

Table 12 indicates that the scores rated by the raters with linguistics or English literature backgrounds were the lowest of the three groups ($M = 4.02$, $SD = 0.75$) while the score rated by the raters with TESOL or ESL education backgrounds were the highest ($M = 4.55$, $SD = 0.70$).

Table 12
Descriptive Statistics of Holistic Scores Based on Raters' Academic Major (N = 62)

Academic Major	*N*	*M*	*SD*
Ling./Literature	9	4.02	.75
TESOL/ESL	38	4.55	.70
Others	15	4.41	.59
Total	62	4.44	.70

To evaluate the relationship between rated holistic scores and the rater's academic major, post hoc multiple comparisons were conducted by using the LSD method to determine if any of the pair-level differences were significant. The independent variable, the rater factor, included three academic major groups. The dependent variable was the holistic rating scores rated by the raters.

Table 13

Post Hoc Multiple Comparisons for Holistic Rating Scores Based on Raters' Academic Major

(I) Major	(J) Major	Mean Difference (I-J)	Std. Error	*Sig.*
Ling./Literature	TESOL/ESL	-.53	.25	.04*
	Other	-.39	.29	.18
TESOL/ESL	Ling./Literature	.53	.25	.04*
	Other	.14	.21	.51
Other	Ling./Literature	.39	.29	.18
	TESOL/ESL	-.14	.21	.51

Note. $*p < 0.05$

Table 13 reveals that there are statistically significant differences between the group of linguistic/English literature raters and the group with TESOL/ESL education ($p = 0.04 < 0.05$). However, compared with the two English-related major groups, the mean scores rated by the group with the non-related English major did not reach the level of statistically significant differences.

In order to evaluate if the raters' teaching experience affected their holistic rating scores, the responders were divided into three groups based on their years of teaching experience: less-experienced (less than 2 years), experienced (3 to 6 years), and well-experienced (more than 7 years).

Table 14

Descriptive Statistics of Rated Holistic Scores Based on Raters' Teaching Experience

Teaching Experience	*N*	*M*	*SD*
Less-experienced	15	4.25	.72
Experienced	22	4.35	.57
Well-experienced	25	4.64	.76
Total	62	4.44	.70

The descriptive statistics are shown on Table 14. Compared with their means of the three groups, the lowest mean scores of the three groups were rated by the less-experienced teachers ($M = 4.25$, $SD = 0.72$), while the highest mean scores were rated by the well-experienced teachers ($M = 4.64$, $SD = 0.76$). In addition, the mean scores of both the less-experienced group and experienced group were lower than the mean scores of the overall holistic rating scores ($M = 4.44$, $SD = 0.70$).

A one-way ANOVA was conducted to evaluate the relationship between the rated scores and the raters' English teaching experience to Taiwanese students. To see if the ANOVA was significant, the

Tests of Between-Subject Effects were examined (See Table 15).

Table 15
ANOVA Results for Rated Holistic Scores Based on the Effect of Raters' Teaching Experience

Source	*Df*	*SS*	*MS*	*F*	*Sig.*
Between Groups	2	01.72	.86	1.83	.17
Within Groups	59	27.73	0.47	-	-
Total	61	29.45		-	-

As results showed in Table 15, there were no statistically significant differences found between the rater groups of English teaching experience to Taiwanese students in holistic rating scores ($p = 0.17 > 0.05$). Although the data shown in Table 14 seemed to say that the more English teaching experience the raters had, the higher their scores were rated, the result did not reach a statistically significant difference.

Findings from the Interview

The main purpose of the individual interview with the raters is to better understand various raters' scoring results and to uncover the variables that have potentially impacted the scores. The researcher arranged face-to-face interviews, telephone interviews, or on-line interviews with the raters who agreed to participate in an interview. During the interview, the researcher asked the same two questions about the way the raters approached and performed the evaluation of the speech samples, and their opinions regarding the analytic components. For interview question two, since the speech sample three received the largest of standardized deviation (*SD* = 1.13, see Table 6) among the five speech samples from the raters, the researcher chose speech sample three to justify why the raters chose the answer they did. The following section presents the abstracts from the raters' interview data.

Interview Question One: When you assess the oral proficiency of the nonnative speakers, especially the Taiwanese college students, what analytic components are you concerned with the most and the least with? Why?

"I concern the most with the speaker's *comprehensibility* to see if he could express his idea clearly and organize words neatly. If he had a hard time to put his thinking meaningfully, I am afraid that he cannot get the high score."

"The part of *grammatical accuracy* affected my rating the most, since those students have learned English for many years. Fluency and pronunciation are difficult to reach at a high level since EFL students do not have enough opportunities to practice these skills with the native speakers. However, grammatical rules are the part they can self-discipline."

"A student's *vocabulary* size can tell me what language proficiency level he is. If he can only use a limited vocabulary, I feel that he is hardly expressing himself."

"*Grammatical accuracy* makes it easy to predict a student's language proficiency level. Without using correct grammatical rules, his speech will confuse his listener or cause misunderstanding.

Grammar is very fundamental for the new language learning."
"As a native speaker, I pay attention to *comprehensibility* to see if his speech makes sense to me. If I could understand what he says without a misunderstanding, he definitely did a good job."
Interview Question Two: This is your holistic and analytic ratings for speech sample three. Can you talk about how you approached the holistic and analytic ratings?
"I noticed that his English was fluent and clear in pronunciation, and I have no trouble with his comprehensibility. However, his speech was too short to use more vocabulary."
"His speech was very clear, and he used correct past-tense verbs all the way. I like the way he spoke since his overall organization was also neat, that's why I gave him a high score."
"He was good at using transition words, and his flow of speech was quite well. However, for the task of picture description, I expected him to describe more detail from the pictures." "Holistically his performance was above average. He had strong grammar and a smooth tone plus he was understandable. However, he should not rush to make a conclusion about his speech."

Discussion and Conclusion

Pedagogical Implications of the Study

This study tried to explore some issues with regard to using the rating scales in foreign language speaking assessment. From the perspective of practical issues, Weigle (2002) stated that a holistic scale rating was a relatively easier and quicker way to assess students' oral skills, while an analytic scale rating was more time-consuming and complicated. However, from the perspective of *washback* issues, a multiple analytic scoring format is more informative than a single holistic scoring format (Nakamura, 2004).

The findings of the present study can enhance the knowledge of different types of rating scales for assessing EFL oral language proficiency, and the relationship between test scores and score meanings: *what* and *how* teachers are concerned with learners' oral performances in L2. Based on the research findings, some pedagogical implications of the study are discussed.

To assess Taiwanese EFL learners' oral proficiency, the rating scales are recommended to use for both native English teachers and nonnative English teachers. The reasons include:

- The functions of the rating scales for testing: it guides the language teacher to select the appropriate tasks for the students, guides the teacher to score the samples, reminds the teacher of the scales/criteria/standards to follow, and maintains the intra-rater reliability and validity of the test.
- Different types of rating scales have their own purposes and characteristics. To choose an appropriate type of rating scale depends on the teacher's particular needs. For instance, a holistic rating scale may be appropriate for placement tests since it can tell the students' overall language proficiency, while an analytic rating scale can be used for assessing the advanced level students' oral performance since it can tell the students' oral skills in each individual component.

- As mentioned in the review of literature, speaking tests are a valuable teaching device for language teachers in the EFL classroom: teachers can receive feedback immediately through their students' performances, and they can give feedback to students based on the descriptor of the rating scales.

According to the findings of the present study, there were two-thirds of the English teachers who had never been trained in rating EFL oral proficiency. Review of related literature indicated that rater training was unable to remove the judges' individual variations or eliminate individual bias; however, the rating training for speaking assessment do benefit the teachers in certain aspects. The training sessions not only can enhance teachers' knowledge of testing speaking abilities and understanding the rating process, but also maintain the teacher's (rater's) judgments to be reliable and consistent.

In the EFL classroom, especially the environment where learners share the same L1 and have limited opportunity for real L2 interactions, Mangubhai (2005) stated that teachers should maximize the target language input to their students, as well as make their classes rich with *comprehensible input* in order to achieve a better language outcome. In addition, with regard to the development of the learners' oral skills, Luchini (2004) suggested that EFL classrooms should create opportunities for learners to participate in an integrating way—both form- and accuracy-focused activities of instruction, since both are believed to contribute to foreign language acquisition.

In conclusion, one of the toughest challenges of oral proficiency testing has been the construction of practical, reliable, and valid tests of oral production ability. Based on the research findings, teachers need to have assistance and encouragement in trying communicative speaking assessment. The ideal test of oral proficiency will be suggested here that it should involve: 1) *live* performance, 2) a careful specification of tasks to be accomplished during the test, and 3) a clear scoring rubric that is truly descriptive of ability.

Finally, teachers should always consider *positive washback*—thebenefit that tests offer to learning—tests therefore will be learning devices through which students can receive a diagnosis of areas of strength and weakness, as well as have clear study goals. For instance, one way to enhance *washback* is to provide "descriptive evaluations" of test performance. Tests therefore will be effective learning devices through which learners can receive a diagnosis of areas of strength and weakness, thus having clear study goals.

References

Bachman, L. F. (1981). An experiment in a picture-stimuli procedure for testing oral communication. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 140-148). Washington, D.C: TESOL.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2),239-257.

Bailey, K. M. (2005). *Practical English language teaching:*

Speaking. New York, NY: McGraw-Hill.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, A. (2004). Discourse analysis and the oral interview: Competence or performance? In D. Box & A. D. Cohen (Eds.), *Studying speaking to inform second language learning* (pp. 253-282). Buffalo: University of Toronto Press.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.
Caban, H. L. (2003). Rater group bias in the speaking assessment of four Japanese ESL students. *Second Language Studies*, *21*(2), 1-44.

Canale, M., & Swain, M. (1980). *Theoretical bases of communicative approaches to second language teaching and testing*. Oxford: Oxford University Press.

Campbell, R. & Wales, R. (1970). The study of language acquisition. In J. Lyons (Ed.), *New horizons in linguistics* (pp. 242-260). Harmondsworth: Penguin Books.

Chen, T. M. (2001). *A study of the relationship between communication strategies and English proficiency evaluation by SLOPE*. Unpublished master thesis, Providence University, Taiwan, R.O.C.

Cheng, L. (2006, May 10). New challenge of new TOEFL. *Lihpao Daily*. Retrieved November 24, 2006, from http://publish.lihpao.com/Education/2006/05/10/06d05103/&w

Davies, A., et al. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Faerch, C., Hasstrup, K., & Phillipson, R. (1984). *Learner language and language learning*. Clevedon, Avon: Multilingual Matters.

Fulcher, G. (2003). *Testing second language speaking*. Great Britain: Pearson Education Limited.

Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.

Huang, Y. T. (2004). Effects of words and pictures on oral performance. *The Proceeding of the 13th Conference on English Teaching and Learning in the Public of China* (pp. 112-122). Taipei: Crane Publishing Co., Ltd.

Hymes, D. (1974). *Foundations of Sociolinguistics: An Ethnographic Approach.* Philadelphia: University of Pennsylvania Press.

Kim, H. J. (2005). *World Englishes and language testing: The influence of rater variability in the assessment process of English language oral proficiency*. Unpublished doctoral dissertation, the University of Iowa.

Kitao, S. K., & Kitao, K. (1996, May). Testing communicative competence. *The Internet TESL Journal, 2(5)*. Retrieved July 21, 2007 from http://iteslj.org/Articles/Kitao-Testing.html

Li, M. C. (2003). *Factors affecting adult EFL learners' speaking performance: Planning and proficiency.* Unpublished master thesis, the National Taiwan Normal University, Taiwan, R.O.C.

Lin, L. T. (1996). *Developing the English oral proficiency in military academic: Its theory and practice.* Unpublished master thesis, the National Political Military University, Taiwan, R.O.C.

Liu, Y. L. (2006). *The effects of rater-related variables on testing oral language ability and assessment of speaking performance.* Unpublished doctoral dissertation, Texas A&M University-Kingsville.

Luchini, P. L. (2004). Developing oral skills by combining fluency-with accuracy-focused tasks: A case study in China. *The Asian EFL Journal, 6*(4). Retrieved March 12, 2007 from http://www.asian-efl-journal.com/december_04_PL.php

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: Cambridge University Press.

Mangubhai, F. (2005). What can EFL Teachers Learn from Immersion Language Teaching? *The Asian EFL Journal*, *7*(4), 203-212.

Na, Z. (2007). A Study of High School Students' English Learning Anxiety. *The Asian EFL Journal, 9*(3). Retrieved November 5, 2007 from http://www.asian-efl-journal.com/Sept_2007_zn.php

Nakamura, Y. (2004). A comparison of holistic and analytic scoring methods in the assessment of writing. *Paper presented at the 3rd annual JALT Pan-SIG Conference*, Tokyo Keizai University, Japan.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York, NY: Peter Lang.

Pan, H. E. (2002). *Motivating beginning EFL learners to speak English in class: An action research in an English speaking class at a junior high school.* Unpublished master thesis, the National Taiwan Normal University, Taiwan, R.O.C.

Savignon, S. J. (1986). Communicative Language Teaching**.** *Theory into Practice*, *26*(4), pp. 235-242.

Shumin, K. (2002). Factors to consider: Developing adult EFL students' speaking abilities. In J. C. Richards, & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 204-211). Cambridge: Cambridge University.

Teng, K. H. (2005). *Perceptions of Taiwanese students to English learning as functions of self-efficacy, motivation, learning activities and self-directed learning.* Unpublished doctoral dissertation, the University of Idaho.

Tsai, C. I. (2003). *Anxiety and beliefs about language learning: a study of Taiwanese college students learning English***.**Unpublished doctoral dissertation, Taxes A&M University-Kingsville, U.S.A.

Ur, P. (1996). *A course in language teaching: Practice and theory*. Cambridge: Cambridge University Press.

Wang, Y. C. (2003). Communication-orientation in Freshman English curriculum: A New response. *The Proceeding of the 12th Conference on English Teaching and Learning in the Public of China* (pp. 588-598). Taipei: Crane Publishing Co., Ltd.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.