English Teaching, Vol. 61, No. 3, Autumn 2006

Rater Reliability in L2 Oral Proficiency Tests

Hyun-Ju Kim
(Dankook University)

Kim, Hyun-Ju. (2006). Rater reliability in L2 oral proficiency tests. *English Teaching, 61*(3), 105-118.

This study attempted to determine (1) whether raters having different English language backgrounds reach similar conclusions regarding Korean students' English language oral proficiency; (2) whether six analytic rating scales of different aspects of proficiency are actually used similarly to the holistic rating scale by various raters having different English language backgrounds. It was found that raters' English language backgrounds influenced their rating behaviors in terms of the analytic rating scales. In addition, NSs in the US showed statistically substantial correlation between holistic and analytic ratings for grammar and pronunciation while on the whole NNSs in Korea showed relationships between holistic and each of the five analytic ratings for grammar, vocabulary, pronunciation, rate of speech, and organization. In terms of inter-rater reliability, the findings imply that non-native speakers of English could be as qualified to assess non-native speakers' English language oral proficiency as native speakers are if they were trained well enough to establish high reliability.

I. INTRODUCTION

Testing English language oral proficiency has been of increasing importance in the field of second language learning and applied linguistics, and it has often been recognized that the best way to test non-native speakers' English language oral proficiency is to measure whether or not they speak like a native speaker. However, many researchers in the field of World Englishes (WEs) have tackled the problems of *teaching* English to non-native speakers and *assessing* their English language proficiency based only on native speakers' norms (Lowenberg, 2002). Measuring non-native speakers' English language oral proficiency is especially critical since more and more localized and nativized Englishes are often used for international communication around the world. However, there has been little attention paid to the way to make it different to measure non-native speakers' English language proficiency and to make the test scores reliable and valid. Along with this issue, this study examines whether raters working independently are apt to reach similar conclusions regarding the English language oral proficiency of Korean students and to what extent

raters having different English language backgrounds are reliable when they evaluate Korean students' English language oral proficiency. The research questions are as follows:

1. Is there a significant difference between native and non-native speakers of English when they evaluate six Korean students' Test of Spoken English (TSE) picture-description task using holistic and analytic rating scales?
2. Is each analytic rating scale weighted similarly in the process of holistic rating of Korean students' TSE picture-description task?

II. LITERATURE REVIEW

Since communicative competence has been of special interest in the field of second language learning and language testing, it has generally been recognized that the best way of measuring communicative competence is testing language oral proficiency. As a result, testing English language oral proficiency has become a more important part of many language proficiency tests such as Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS). However, the issue of who, what, and how to test non-native speaker's English language oral proficiency remains problematic.

English language oral proficiency is usually evaluated by human raters, mostly native speakers (Brown, 2004). Raters play a major role in the assessment process and influence the quality and meaning of the scores obtained. Douglas (1997) states:

> To attempt to isolate any single component of language ability may be fruitless. We need to know more about how raters arrive at judgments, what aspects of the discourse they attend to in making their ratings, and how different raters arrive at similar ratings for perhaps very different reasons (p. 22).

There have been some previous studies of the relationship between raters and test scores of L2 oral performance assessment (Bachman et al., 1995; Brown, 1995; Lumley & McNamara, 1995; Lynch & McNamara, 1998; Upshur & Turner, 1999; Weigle, 1998). All these studies have found significant differences among raters. Brown (1995) and Hill (1996) have assumed that different ratings can be controlled for by rater training with explicit assessment criteria and samples of performance at different levels.

However, there have been two opposing literatures on rater training. Some researchers (Diederich et al., 1961; Shohamy et al., 1992) argue that rater training has an important effect on minimizing inconsistent ratings. Others (Lumley & McNamara, 1995; Lunz et al., 1990; Weigle, 1998) argue that rater training does not have a significant effect on accurate

or consistent ratings for the long-run. In the measurement literature it is argued that rater variation is an inevitable part of the rating process and it should be considered as a benefit because it provides enough variability to allow probabilistic estimation of rater severity, task difficulty, and test-taker's ability (Weigle, 1998).

Test scores can be interpreted or even measured in a different manner according to different techniques for eliciting test scores (Madeson, 1980). Two ways of assessing students' speech, using holistic and analytic scales, have been well documented (Bachman, 1988; Bachman & Savignon, 1986; Douglas & Smith, 1997; Fulcher, 1997; Ingram & Wylie, 1993; Weir, 1990). Holistic scoring concentrates on communication and requires that raters evaluate a wide variety of criteria simultaneously. Although holistic rating is desirable for the evaluation of the overall communicative effectiveness of the test-taker (Weir, 1990), raters can be confused when evaluating many things simultaneously (Bachman, 1990). On the other hand, analytic scoring procedures attempt to identify smaller units such as grammar, vocabulary, pronunciation, and organization. The proponents of analytic rating scales (Bachman & Savignon, 1986) argue that since speaking ability is a multi-componential trait, rating scales should be defined in terms of components such as functional, grammatical, discourse, and sociolinguistic competence to reflect the Communicative Language Ability (CLA) model. However, if a rater fails to use the rating scales appropriately, the test scores will eventually be affected (Upshur & Turner, 1999). Therefore, as Bachman and Savignon (1986) suggested, holistic ratings along with analytic ratings should be assigned to provide a precise profile of the test-taker's speaking ability.

III. METHODS

This study is designed to determine if raters having different English language backgrounds reach the same judgments using holistic and analytic rating scales when they assess Korean students' English language oral proficiency. In addition, this study is designed to determine the relative weight given to each component category in predicting the overall proficiency score. To answer the research questions, a single-factor experimental design having repeated measure was chosen. All tests were performed using the Statistical Package for the Social Sciences (SPSS) for Windows version 12.0.

1. Participants

The participants consisted of two groups of raters: 30 native speakers in the US and 30 non-native speakers in Korea. The US represents the context of English as a Native Language (ENL) and Korea represents that of English as a Foreign Language (EFL). All

participants had taught English for at least two years and had been instructed on how to use the rating instruments with an instructional letter.

2. Speech Samples

I used six TSE speech tapes at various proficiency levels provided by the Educational Testing Service (ETS). All of the examinees who provided speech samples were Korean test-takers and I selected a picture-description task for use in this study. Although the test scores of this particular picture-description task can be influenced by test-takers' cultural backgrounds (Byram, 1997) and test scores can vary from task to task (Chalhoub-Deville, 1995a, 1995b, 2001; Derwing et al., 2004; Foster & Skehan, 1996), there are several reasons that I selected only one picture-description narrative task. First, the narrative task has been used in many studies and it is not influenced by interactional variables as a dialogue task is (Yuan & Ellis, 2003). Second, the one narrative task is not so demanding for raters to evaluate since it does not require them to evaluate test-takers' cognitively complex processes (Brown et al., 2005).

3. Rating Instruments

Raters were instructed to rate the speech sample based on their overall impression of each speaker on a 7-point Likert scale with 1 representing low proficiency, 4 representing moderate proficiency, and 7 representing high proficiency. Each rater assessed the same speech samples twice: using a holistic rating scale for overall impression and using analytic rating scales for six rating components – grammatical accuracy, vocabulary, pronunciation/ accent, rate of speech, organization, and task fulfillment. In terms of the analytic rating scales, I selected them based on the results of the survey of rating instruments of international English language oral proficiency.

4. Procedures

In order to assess the six speech samples, the raters were first given ten minutes to overview the analytic rating criteria and then they provided ratings on analytic scales encompassing grammatical accuracy, vocabulary, pronunciation/accent, rate of speech, organization, and task fulfillment. Raters listened to the one-minute speech of English again and provided holistic scores reflecting their overall impression of the level of English language oral proficiency to each of the six speech samples within 20 seconds. I administered this procedure separately for each group (native speakers in the US and non-native speakers in Korea).

5. Research Design and Analysis

To examine the research questions, I used the repeated measure of one-way analysis of variance (ANOVA) technique. All tests were performed using the Statistical Package for the Social Sciences (SPSS) for Windows version 12.0. The index to observe for significance is the ANOVA F-value. All tests were performed at the .05 significance level. For reliability, inter-rater reliabilities were calculated among the 30 native speakers in the US and among the 30 non-native speakers in Korea. In order to investigate the relationship between holistic and analytic ratings, the correlation coefficients were computed in SPSS.

IV. RESULTS

1. Descriptive Statistics of Holistic Ratings

Overall means and standard deviations (SD), along with minimum and maximum holistic ratings from 60 raters for each of the six speech samples on the picture-description task, are presented in Table 1. The lowest rating obtained is 1 and the highest is 7, indicating that raters used the whole range of the rating scale. The means range from 1.92 to 6.77 and the SD range from .50 to 1.32, which indicates that English language oral proficiency test scores of the speech samples varied considerably. As for the differences in raters' holistic mean ratings for each group, the holistic scores of speech sample one ranged from 3.00 to 7.00 with SD .90 by NSs and from 2.00 to 6.00 with SD .93 by NNSs. The holistic scores of speech sample two ranged from 1.00 to 6.00 with SD 1.09 by NSs and from 1.00 to 5.00 with SD .95 by NNSs. The holistic scores of speech sample three ranged from 2.00 to 7.00 with SD 1.32 by NSs and again from 2.00 to 7.00 with SD 1.03 by NNSs. The holistic scores of speech sample four ranged from 2.00 to 7.00 with SD 1.16

TABLE 1
Descriptive Statistics of Holistic Ratings

Rater		A 1	A 2	A 3	A 4	A 5	A 6
NS	Mean	4.87	2.83	4.67	4.37	2.33	6.77
	SD	.90	1.09	1.32	1.16	.99	.50
	Min	3.00	1.00	2.00	2.00	1.00	5.00
	Max	7.00	6.00	7.00	7.00	4.00	7.00
NNS	Mean	4.13	2.45	5.03	4.05	1.92	6.58
	SD	.93	.95	1.03	1.06	.71	.50
	Min	2.00	1.00	2.00	2.00	1.00	6.00
	Max	6.00	5.00	7.00	6.00	4.00	7.00

Note: in the designation, A indicates speech samples.

by NSs and from 2.00 to 6.00 with SD 1.06 by NNSs. The holistic scores of speech sample five ranged from 1.00 to 4.00 with SD .99 by NSs and from 1.00 to 4.00 with SD .71 by NNSs. Lastly, the holistic scores of speech sample six ranged from 5.00 to 7.00 with SD .50 by NSs and from 6.00 to 7.00 with SD .50 by NNSs. Speech sample five received the lowest mean ratings (2.33 by NSs and 1.92 by NNSs) and speech sample six received the highest mean ratings (6.77 by NSs and 6.48 by NNSs).

2. Inferential Statistics of Holistic and Analytic Ratings

Tables 2 through 8 show the results of one-way analysis of variance (ANOVA) having repeated measure on speech samples. The dependent variables were English language oral proficiency scores reported using holistic and analytic scales. In terms of independent variables, only the significance for the between-subject factor, rater's English language background, will be discussed. This is because in the present study I expected to have significant effects for the speech samples which reflect different proficiency levels. As a result, it was not substantively important to find effects for speech samples on holistic and analytic scores. Therefore, in this section only the significance found in the between-subject factor will be discussed.

TABLE 2
Repeated One-way ANOVA Analyses for Holistic Scores

Source	SS	df	MS	F	P
Between Subjects					
B	6.67	1	6.67	2.94	.09
Subjects (B)	131.49	58	2.27		
Within Subjects					
A	783.78	5	156.76	232.96	.00
A x B	9.58	5	1.92	2.85	.02
A x Subject (B)	195.14	290	.67		

Note: in the designation, A indicates speech samples and B indicates the raters' English language backgrounds.

TABLE 3
Repeated One-way ANOVA Analyses for Grammatical Accuracy

Source	SS	df	MS	F	P
Between Subjects					
B	.71	1	.71	.28	.60
Subjects (B)	149.51	58	2.58		
Within Subjects					
A	477.32	5	95.46	112.74	.00
A x B	9.12	5	1.82	2.16	.06
A x Subject (B)	245.56	290	.85		

TABLE 4
Repeated One-way ANOVA Analyses for Vocabulary

Source	SS	df	MS	F	P
Between Subjects					
B	1.88	1	1.88	.73	.40
Subjects (B)	148.78	58	2.57		
Within Subjects					
A	414.12	5	82.82	114.69	.00
A x B	4.46	5	.89	1.23	.29
A x Subject (B)	209.42	290	.72		

TABLE 5
Repeated One-way ANOVA Analyses for Pronunciation

Source	SS	df	MS	F	P
Between Subjects					
B	2.03	1	2.03	.60	.44
Subjects (B)	197.14	58	3.40		
Within Subjects					
A	431.38	5	86.28	95.41	.00
A x B	5.23	5	1.05	1.16	.33
A x Subject (B)	262.23	290	.90		

TABLE 6
Repeated One-way ANOVA Analyses for Rate of Speech

Source	SS	df	MS	F	P
Between Subjects					
B	11.38	1	11.38	4.22	.04
Subjects (B)	156.46	58	2.70		
Within Subjects					
A	539.00	5	107.80	129.46	.00
A x B	5.19	5	1.04	1.25	.29
A x Subject (B)	241.48	290	.83		

TABLE 7
Repeated One-way ANOVA Analyses for Organization

Source	SS	df	MS	F	P
Between Subjects					
B	57.60	1	57.60	17.60	.00
Subjects (B)	189.83	58	3.24		
Within Subjects					
A	352.37	5	70.47	84.96	.00
A x B	8.73	5	1.75	2.11	.07
A x Subject (B)	240.57	290	.83		

TABLE 8
Repeated One-way ANOVA Analyses for Task Fulfillment

Source	SS	df	MS	F	P
Between Subjects					
B	32.40	1	32.40	7.33	.01
Subjects (B)	256.53	58	4.42		
Within Subjects					
A	381.40	5	76.28	95.76	.00
A x B	14.27	5	2.85	3.58	.00
A x Subject (B)	231.00	290	.80		

The results of the repeated one-way ANOVA test show that there was no significant main effect of raters' English language backgrounds in the holistic ratings. However, there were significant main effects of raters' English language backgrounds in some analytic ratings such as in rate of speech ($F=4.22, p=.04$), organization ($F=17.60, p=.00$), and task fulfillment ($F=7.33, p=.01$). In other words, there were significant differences among the two different rater groups (NSs in the US and NNSs in Korea) when they evaluated Korean students' English language oral proficiency on rate of speech, organization, and task fulfillment using analytic scales although there were no significant differences among these rater groups in holistic ratings and in analytic ratings for grammar, vocabulary, and pronunciation. This fact implies that raters' English language backgrounds influence the analytic ratings for rate of speech, organization, and task fulfillment. However, raters' English language backgrounds does not affect the evaluation of the Korean students' speech samples in holistic and some commonly used analytic rating scales (grammatical accuracy, vocabulary, and pronunciation).

3. Rater Reliability

The reliability estimates indicate the extent of agreement among raters within each of the two rater groups (NSs in the US and NNSs in Korea) on the holistic and analytic ratings of the six speech samples. Table 9 presents the reliability estimates for all the raters averaged together. Table 9 shows that for two different rater groups, the inter-rater reliability indices for the holistic and the six analytic rating scales are statistically substantial although NSs in the US provided higher reliability indices to some extent than NNSs in Korea. Despite no rater training, the statistically substantial reliability indices support the quality of ratings provided by the two different rater groups categorized by their English language backgrounds. It also implies that non-native speakers are also qualified in evaluating non-native speakers' English language oral proficiency and their reliability can be improved with advanced training and rating practices.

TABLE 9
Reliability of Holistic and Analytic Ratings Across the Six Speech Samples

	NSs in the US (n=30)	NNSs in Korea (n=30)
Holistic	.76	.72
Grammatical Accuracy	.75	.72
Vocabulary	.76	.74
Pronunciation	.76	.73
Rate of Speech	.76	.76
Organization	.78	.73
Task Fulfillment	.79	.74

4. Correlations Between Holistic and Each of the Six Analytic Rating Scales

The correlation shows the degree of relationship between two variables. Table 10 shows to what extent holistic rating scores are related to each of the six analytic rating scores provided by the NSs in the US. The intercorrelations between the holistic and six analytic rating scores indicated that the holistic rating scores and the analytic rating scores for grammar and pronunciation measured by the NSs in the US were substantially related. That is, the higher analytic scores for grammar and pronunciation the Korean students' speech samples were given by the NSs in the US, the higher holistic scores the same speech samples were given by the same rater group. As for intercorrelation values, NSs in the US produced the highest intercorrelation value between grammar and vocabulary (.82) while the NSs in the US disagreed in holistic ratings and the other four analytic ratings for vocabulary, rate of speech, organization, and task fulfillment; they somewhat highly agreed in the analytic ratings each other (.47 to .82).

TABLE 10
Inter-correlations for Holistic Scores and Six Analytic Scores Measured by NSs in the US

	Hol	Gram	Voca	Pron	Rate	Org	Task
Hol	--						
Gram	.37*	--					
	(.04)						
Voca	.34	.82**	--				
	(.07)	(.00)					
Pron	.58**	.74**	.70**	--			
	(.00)	(.00)	(.00)				
Rate	.31	.58**	.77**	.70**	--		
	(.10)	(.00)	(.00)	(.00)			
Org	.32	.57**	.78**	.64**	.76**	--	
	(.08)	(.00)	(.00)	(.00)	(.00)		
Task	.26	.47**	.57**	.51**	.54**	.76**	--
	(.16)	(.00)	(.00)	(.00)	(.00)	(.00)	

Note: Hol=holistic, Gram=grammatical accuracy, Voca=vocabulary, Pron=pronunciation, Rate=rate of speech, Org=organization, Task=task fulfillment; the indices in parenthesis indicate significance values.

TABLE 11
Inter-correlations for Holistic Scores and Six Analytic Scores Measured by NNSs in Korea

	Hol	Gram	Voca	Pron	Rate	Org	Task
Hol	--						
Gram	.45** (.00)	--					
Voca	.66** (.00)	.60** (.00)	--				
Pron	.62** (.00)	.39* (.02)	.62** (.00)	--			
Rate	.54** (.00)	.46** (.00)	.62** (.00)	.62** (.00)	--		
Org	.59** (.00)	.56** (.00)	.63** (.00)	.47** (.00)	.62** (.00)	--	
Task	.30 (.06)	.50** (.00)	.32 (.05)	.30 (.07)	.35* (.03)	.74** (.00)	--

Table 11 shows the intercorrelations between the holistic and six analytic rating scores when the NNSs in Korea evaluated the Korean students' speech samples. It indicates that the holistic rating scores and each of the analytic rating scores for grammar (.45), vocabulary (.66), pronunciation (.62), rate of speech (.54), and organization (.59) were positively related. That is, the higher holistic scores the Korean students' speech samples were given by the NNSs in Korea, the higher analytic scores the same speech samples were given by the same rater group. It seemed that the evaluation of the Korean students' English language oral proficiency by the NNSs in Korea was more appropriate than that by NSs in the US since the holistic ratings given by the NNSs in Korea were generally related to most analytic rating scales. Based on the intercorrelation values (.30 to .74), NNSs in Korea produced the highest intercorrelation value between organization and task fulfillment (.74). These correlation results indicate that NSs and NNSs groups in this study have used the rating scales differently for assessing Korean students' English language oral proficiency.

V. CONCLUSION AND DISCUSSION

All tests are subjective to some degree in designing and item writing. Therefore, objective tests are generally considered in terms of score marking. Subjectivity usually extends to the score marking in speaking tests due to human ratings. Since the ratings of non-native speakers' English language oral proficiency generally reflect raters' assessment schemes in addition to test-takers' true abilities, and since the rater groups with different English language backgrounds evaluate English language oral proficiency differently (Bachman et al., 1995; Brown, 1995; Chalhoub-Deville, 1995a, 1995b; Douglas, 1997;

Hill, 1996; Lumley & McNamara, 1995; Lynch & McNamara, 1998; O'Loughlin, 2002; Upshur & Turner, 1999; Weigle, 1998), the rater has remained as an important factor influencing the validity as well as the reliability of test scores.

In line with previous research, this study investigated whether raters having different English language backgrounds reach similar conclusions when they evaluate Korean students' English language oral proficiency. I examined not only how NSs in the US and NNSs in Korea holistically evaluated the speech samples, but also how they evaluated the same speech samples analytically. The results indicate that NSs and NNSs groups generally provided similar overall scores for the Korean students' English language oral proficiency since no significant differences were found in the holistic scoring of the six Korean students' speech samples. However, the two rater groups differed in some analytic ratings of the Korean students' speech samples such as in rate of speech, organization, and task fulfillment. These findings have implications for controlling rater variables such as raters' English language backgrounds.

In addition, for these two rater groups with different English language backgrounds, the inter-rater reliability for the holistic and the six analytic rating scales was quite reasonable statistically on the condition of no rater training in this study. The reasonable reliability not only supports the findings that the ratings provided by the two different rater groups categorized by their English language backgrounds are valuable in this study, but also implies that both the NSs in the US and the NNSs in Korea are qualified as the raters for assessing Korean students' English language oral proficiency.

One of the challenges that NNSs in Korea face is that they, as non-native speakers of English, are merely assumed not to be qualified to teach English through English or to measure Korean students' English performance; moreover, only NSs from western countries are assumed to be well-qualified to teach and test the English language reliably. However, the results in this study refute those prejudices; instead, the results support the credibility of non-native English teachers' assessment of non-native speakers' English language oral proficiency with the statistically reasonable indices of inter-rater reliability within the group of NNSs in Korea.

For the rating behaviors with the holistic and analytic rating scales by the two different rater groups, it was found that the holistic ratings of Korean students' speech samples by NSs in the US were related more to the analytic ratings for grammar and vocabulary than other analytic ratings for pronunciation, rate of speech, organization, and task fulfillment. On the other hand, the holistic ratings by NNSs in Korea were not related to a specific analytic rating scale, but rather their holistic ratings were related to all of the analytic ratings except the analytic ratings for task fulfillment. Therefore, it implied that NNSs in Korea did not make a decision of the holistic test scores by some specific analytic rating components. For the NSs in the US, they seemed to decide the holistic test scores of the Korean students' speech samples, related with some analytic rating scores such as grammar and pronunciation.

The issue of rater inconsistency has always been a concern in the language testing field. In order to improve the validity of test score interpretation, raters' reliability in L2 oral tests is the one of the important ongoing concern among the researchers and testers of English language oral proficiency. This research demonstrates not only that raters' English language backgrounds affect some analytic rating scores, but also that the ratings of non-native speakers' English language oral proficiency by NNSs in Korea are reliable as those by NSs in the US. Finally, this study proposes that the assessment of non-native speakers' English language oral proficiency could be done by non-native speakers of English as well as by native speakers if they are trained well with appropriate rating scales.

REFERENCES

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149-164.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

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*, 238-252.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380-390.

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., Iwashita N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *TOEFL Monograph Series, 29*, Educational Testing Service.

Brown, J. B. (2004). What do we mean by bias, Englishes, Englishes in testing, and English language proficiency? *World Englishes, 23*(2), 317-319.

Byram, M. (1997). *Teaching and assessing intercultural communicative competence.* Clevedon: Multilingual Matters.

Chalhoub-Deville, M. (1995a). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-33.

Chalhoub-Deville, M. (1995b). A contextualized approach to describing oral language proficiency. *Language Learning, 45*, 251-281.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Testing, 54*(4), 655-679.

Diederich, P. B., French, J. E., & Carlton, S. T. (1961). Factors in judgments of writing ability. *Research Bulletin* 61-15. Princeton, NJ: Educational Testing Service.

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. *TOEFL Monograph Series, Number 8*. Princeton: Educational Testing Service.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project.* TOEFL Monograph Series No. 9. Princeton, NJ: Educational Testing Service.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition, 18*(3), 299-324.

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham & D. Corson (Eds.), *The encyclopedia of language education: Vol. 7. Language testing and assessment* (pp. 75-85). Dordrecht, Netherlands: Kluwer Academic.

Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing, 5*(2), 29-50.

Ingram, D., & Wylie, E. (1993). Assessing speaking proficiency in the International English Language Testing System. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium* (pp. 220-234). Alexandria, Virginia: TESOL, Inc.

Lowenberg, P. (2002). Assessing English proficiency in the expanding circle. *World Englishes, 21*(3), 431-435.

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

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Madeson, H. S. (1980). Selecting appropriate elicitation techniques for oral proficiency tests. In A. S. John (Ed.), *Directions in language testing* (pp. 87-99). Singapore: Singapore University Press.

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

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*, 26-33.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111.

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

Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*(1), 1-27.

Applicable levels: tertiary and adult education
Key words: reliability, English language oral proficiency, holistic rating, analytic rating

Hyun-Ju Kim, Ph.D
Dept. of English Language and Literature
Dankook University
147 Hannam-dong, Yongsan-gu
Seoul, Korea. 140-714
E-mail: hyunjukim@dankook.ac.kr

Received in May, 2006
Reviewed in June, 2006
Revised version received in August, 2006