영어교육, 60권 4호 2005년 겨울

Challenge of World Englishes to Language Testing: Investigation of Rater Variability in the Assessment Process

Hyun-Ju Kim
(Hanyang Cyber University)

Kim, Hyun-Ju. (2005). Challenge of world Englishes to language testing: Investigation of rater variability in the assessment process. *English Teaching, 60*(4), 533-548.

This study examined whether or not there was an interaction between raters' English language backgrounds and their attitudes toward World Englishes (WEs) in their rating of Korean students' English language oral proficiency on the Test of Spoken English (TSE) picture-description task using holistic and six analytic rating scales (grammatical accuracy, vocabulary, pronunciation/accents, rate of speech, organization, and task fulfillment). This study also examined whether or not raters' English language backgrounds and their attitudes toward WEs in language testing affected their rating of six Korean students' speech samples using holistic and analytic rating scales. It was found that raters had substantial variability in the process of assessing the speech samples when they use analytic rating scales and these raters' ratings were revealed to be influenced more by their attitudes toward WEs in language testing than by their English language backgrounds.

I. INTRODUCTION

One of the most important issues in English Language Teaching (ELT) at present is the recognition that English is increasingly used for international communication (McKay, 2002). If one is learning English in an English as a Foreign Language (EFL) context, it becomes necessary to learn English as an international or a world language rather than as a foreign language: today, people in EFL countries are using English within the country as well as for international communication (Kachru, 1994; Seidlhofer, 1999). International communication takes place not just with native speakers but also with anyone who speaks English outside of their own country. Native speakers are actually a minority of English speakers in the world, as there are far more non-native speakers of English in the world than there are native speakers (Kachru, 1997;

Pennycook, 1999, 2001). In addition, the native speaker's standard or 'correct' English in terms of grammar and phonology in the native country is not always regarded as useful or appropriate in international contexts.

It has been said that the goal in language assessment is to reduce sources of error that are external to the learner's language performance to the greatest possible degree in order to reflect the candidate's true ability (Wigglesworth, 2001, p. 188). The sources of error in oral test performance are varied. Among the various sources of error, many researchers (Brown, 1995; Lumley & McNamara, 1995; Lunz et al., 1990; Lynch & McNamara, 1998; Weigle, 1994; Wigglesworth, 1993) have studied rater effect on test scores. However, there is no research on rater effect on test scores in the WEs perspective. In order to examine the variability in the rating process and product across different groups of raters, research is needed to study the ratings of non-native speakers' English language oral proficiency provided by non-native speakers as well as by native speakers. This study will address the following research questions:

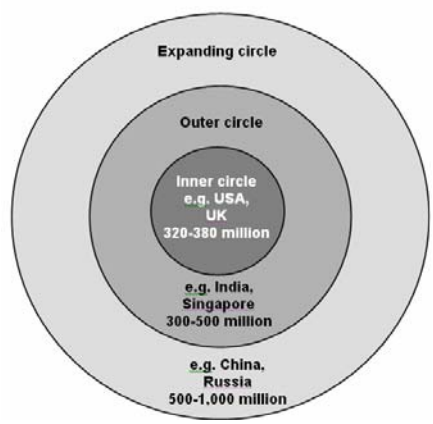
1. Is there a significant interaction between raters' English language backgrounds and their attitudes toward WEs for language testing in the holistic and analytic ratings of Korean students' speech samples on the TSE picture-description task?
2. Is there a significant difference among different groups of raters (native speakers in the US, native speakers in Korea, non-native speakers in Hong Kong, and non-native speakers in Korea) in the rating of Korean students' speech samples on the TSE picture-description task using holistic and analytic scores as measures?
3. Is there a significant difference among different groups of raters based on their attitudes (negative, neutral, and positive) toward WEs for language testing in the rating of Korean students' speech samples on the TSE picture-description task using holistic and analytic scores as measures?

II. THEORETICAL BACKGROUND

1. World Englishes

Kachru (1982) proposed the three-concentric circle model of WEs. His model has been widely used as a standard framework for WEs studies (see Figure 1). This model categorizes English speakers into three groups: the inner circle, the outer circle, and the expanding circle. The inner circle contains English as a Native Language (ENL) speakers and includes countries such as the UK, the USA, Ireland, Canada, Australia, and New Zealand. The outer circle contains English as a Second or Additional Language (ESL/EAL)

FIGURE 1
Three Concentric Circles of Englishes



speakers, which is used for intra-national communication and includes countries such as India, Nigeria, Hong Kong, Singapore, and the Philippines. The expanding circle contains English as a Foreign Language (EFL) speakers, and recognizes the importance of English as an international language. It includes countries such as China, Russia, Brazil, Korea, Japan, and Taiwan. This model also shows the global situation of English. That is, the inner circle comprises the longstanding English-using countries, the outer circle comprises the institutionalized English-using countries, and the expanding circle comprises the English-using countries in which English has various roles and is used for more limited purposes than in the outer circle. Actually, most learners of English in the expanding circle countries rarely have much contact with native speakers (Kachru & Nelson, 2001).

Kachru's models of WEs (1985, 1997), which distinguished English speakers by geographic locations and historical reasons, have been criticized by several researchers (Jenkins, 2003; Modiano, 1999a, 1999b) in that there are some countries in the expanding circle that have their localized form of English that use it as the norm in their communities (Shim, 1999). McArthur (1987, p. 11) proposed a model of WEs that highlights a broad spectrum ranging from 'world standard English' to the various national and regional Englishes. More recently, Modiano (1999a) proposed that the inner circle be occupied by proficient speakers of international English, which includes all of the varieties of English used in cross-cultural communication. Modiano (1999a, p. 26) argued that "as the need to communicate internationally increases, more and more people of this second circle move inward and in time obtain levels of proficiency which place them in the first circle." His model is very compelling in that it breaks up the notion of the importance of native-like proficiency and it does not require native-like accents or lexical registers for effective international communication.

However, the problem with this model is that it is not clear who decides what constitutes "proficiency" and what it means to be proficient in international English (Modiano, 1999b).

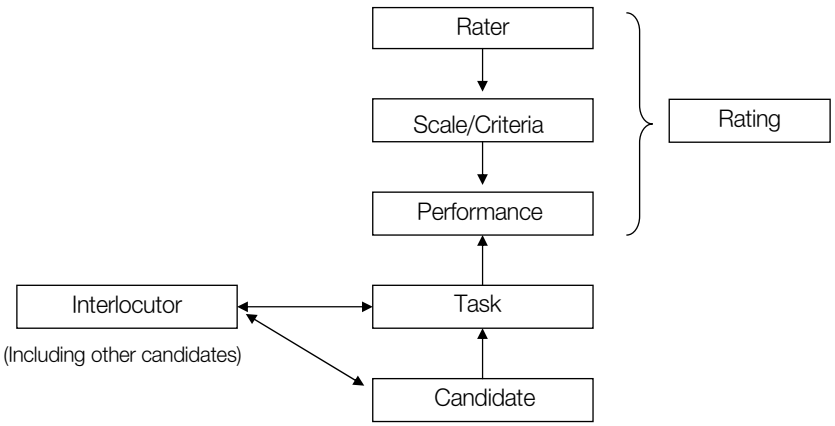
In Modiano's (1999b) revised model, he argued that there is a "common core" in every variety of English, which he called English as an International Language (EIL), and anyone who is proficient in international English is included in the inner circle. Standards are not solely dependent on native speakers' norms. His revised model still has limitations: it is still uncertain what EIL means and who should evaluates the proficiency levels of speakers of EIL. Yano (2001) also argued that as more English varieties become firmly established, the necessity of seeking correct models in varieties of Englishes spoken by the genetically-native speakers will decline. In sum, WEs models describe the reason why the varieties of Englishes are labeled WEs and propose that researchers in language learning and testing pay attention to the changing uses of English in various contexts and application of the WEs perspective into language testing.

2. L2 Oral Test Performance

The rules of speaking continually change with time and place (Kachru & Nelson, 2001). Although Bachman's (1990) "Communicative Language Ability (CLA)" model has been regarded as the best depiction of language test performance, several researchers have raised questions about other factors influencing test scores (Chalhoub-Deville, 2003; Chalhoub-Deville & Deville, 2005; McNamara, 2003).

There are some missing factors connected to the test scores, but there has not been a comprehensive understanding of these factors in a language testing performance yet (McNamara, 1996, p. 85). McNamara pointed out that it was necessary to consider 'rating' as a factor that affects language performance and test scores in language testing. He said that 'rating is a result of a host of factors interacting with each other' (McNamara, 1997, p. 453). McNamara broadened the meaning of 'interaction' in language testing and included the process of rating. He interpreted the rating as an end-product after all of the interaction process happens among task, test-taker, testing performance, rating criteria, rater, and interlocutor. That is, he argued that the 'interpretation of performance is inherently a social act.' The interaction in performance assessment of speaking skills was described as follows:

FIGURE 2
'Proficiency' and Its Relation to Performance (McNamara, 1996, p. 86)



This model should be adopted in the validation process of interpretation of test scores of English language oral proficiency tests. The effect of the potential variables such as rater, rating criteria, and task on test scores should be thoroughly researched to be included as an important part of an assessment framework. Since the rater is the one of the important sources that influences test scores, it is especially necessary to study in which way and to what extent rater characteristics affect their rating of non-native speakers' English language oral proficiency. Test scores are closely linked to tasks, raters, and rating scales. Relating to the theory of WEs, this study explores rater variable. The present study investigates whether or not raters' English language backgrounds and raters' attitudes toward WEs in language testing affect their ratings of non-native speakers' English language oral proficiency.

III. METHOD

1. Participants

The participants consisted of four groups of raters: 30 native English language teachers in the US, 35 native English language teachers in Korea, 30 non-native English language teachers in Hong Kong, and 38 non-native English language teachers in Korea. The countries of the US, Hong Kong, and Korea were chosen for this study because each of them represents English as a Native Language (ENL), English as a Second Language (ESL), and English as a Foreign Language (EFL) context. 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. Instrument

Test scores can be 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, were used in this study. Each rater assessed the same six speech samples twice: first using a holistic scale for overall impression and then using analytic scales for six rating components – grammatical accuracy, vocabulary, pronunciation/ accent, rate of speech, organization, and task fulfillment - on a 7-point Likert scale with 1 representing low proficiency and 7 representing high proficiency. Since speaking ability is a multi-componential trait (Bachman & Savignon, 1986), rating scales were defined in terms of components such as functional, grammatical, discourse, and sociolinguistic competencies to reflect the Communicative Language Ability (CLA) model.

3. Procedure

I sent participants a package consisting of questionnaires, CDs with the six speech samples on the Test of Spoken English (TSE) picture-description task, a letter to raters that gave instructions for rating speech samples, a rating booklet which included the analytic rating scale descriptors as well as holistic and analytic scoring forms, and a consent form. To minimize the "halo effect[1]" in rating, I asked the participants to (1) evaluate the six speech samples on one CD using holistic scales, (2) complete the questionnaire, (3) review the analytic rating scale descriptors, and then (4) evaluate the same six speech samples, ordered differently on a second CD, using analytic scales. After completing the questionnaire and evaluation of the six speech samples, raters were asked to send their completed questionnaires and rating instruments to me.

I used SPSS (Statistical Package for the Social Sciences) 12.0 for Windows and employed repeated measures of three-way analysis of variance (ANOVA) with two between-subjects factors (raters' English language backgrounds and their attitudes toward World Englishes in language testing) and a single within-subject factor of six levels (speech samples). The dependent variables were English language oral proficiency scores measured using holistic and analytic rating scales.

[1] Halo effect means that when raters consider a person good (or bad) in one category, they are likely to make a similar evaluation in other categories (Thorndike, 1920).

IV. RESULTS

In order to identify raters' attitudes toward WEs in language testing, I decided to divide the raters into three WEs attitude (negative, neutral, and positive) groups: approximately one third of the total raters were positive raters, one third of total raters were neutral raters, and one third of total raters were negative raters. There were four raters who had 34 total score and five raters who had 35 total score on the questionnaire about the WEs attitude. In other words, raters' total scores were tied in cut-off scores when the raters were divided into three rater groups evenly in one-third. As a result, if a rater's total score on the attitude questionnaire was in the top 31.6% of all raters' scores, the rater was referred to as positive toward WEs for language testing. A rater who had a total score on the attitude questionnaire in the bottom 33.8% of all raters' scores was referred to as negative toward WEs for testing purposes. Table 1 presents a summary of frequencies and percentages of the participants' responses to the attitude questionnaire.

TABLE 1
Percentage of Raters' Attitudes Toward WEs in Language Testing

Attitudes toward WEs in language testing	Total score range	All raters (n=133)	NSs in the US (n=30)	NSs in Korea (n=35)	NNSs in HK (n=30)	NNSs in Korea (n=38)
Negative	16-34	42 (31.6%)	6 (20.0%)	13 (37.1%)	3 (10.0%)	11 (28.9%)
Neutral	35-42	46 (34.6%)	5 (16.7%)	13 (37.1%)	12 (40.0%)	18 (47.4%)
Positive	43-60	45 (33.8%)	19 (63.3%)	9 (25.7%)	15 (50.0%)	9 (23.7%)

There were more positive raters than neutral or negative ones in the groups of NSs in the US and NNSs in Hong Kong. There were more neutral raters (47.4%) in the group of NNSs in Korea than in the other rater groups. There were more negative raters (37.1%) in the group of NSs in Korea than in other rater groups. To confirm whether or not there is a statistically significant difference among NSs in the US, NSs in Korea, NNSs in Hong Kong, and NNSs in Korea in their attitudes toward WEs in language testing, I employed one-way ANOVA in SPSS. Raters' total scores on the attitudinal items were treated as a dependent variable. Raters' English language backgrounds were treated as an independent variable. Table 2 shows that there is a significant difference among different rater groups in their attitudes toward WEs in language testing.

TABLE 2
Summary of One-way ANOVA for the Attitudes Toward WEs in Language Testing

	SS	df	MS	F	P
Between Groups	1507.74	3	502.58	7.50*	.00
Within Groups	8644.89	129	67.02		
Total	10152.63	132			

* The mean difference is significant at the .05 level.

As shown in Table 2, raters' English language backgrounds significantly affected their attitudes toward WEs in language testing. In order to determine which group means differ significantly from the others, post hoc tests were performed. Results from the Bonferroni test are shown in Table 3. In Table 3, post hoc analyses of the data for the total scores on attitude items of the questionnaire revealed that there are significant mean differences between NSs in the US and NSs in Korea, and between NSs in the US and NNSs in Korea.

TABLE 3
Bonferroni Multiple Comparisons of Attitude Scores of the Questionnaire

English Language Background	English language background	Mean Difference	Std. Error	Sig.
NS in the US	NS in Korea	7.82*	2.04	.00
	NNS in HK	2.43	2.11	.66
	NNS in Korea	7.74*	2.00	.00
NS in Korea	NNS in HK	-5.39	2.04	.06
	NNS in Korea	-.08	1.92	1.00
NNS in HK	NNS in Korea	5.31	2.00	.05

* The mean difference is significant at the .05 level.

In terms of the research questions, the results of the repeated three-way ANOVA showed that there was no interaction between English language backgrounds and attitudes toward WEs in language testing when raters evaluate non-native speakers' English language oral proficiency using holistic and analytic rating scales. There were significant main effects for raters' English language backgrounds in analytic ratings on grammar and organization, but not in holistic ratings (see Tables 4, 5, & 6).

TABLE 4
Summary of the Repeated Three-way ANOVA Using Holistic Rating Scales

Source	SS	df	MS	F	P
Between Subjects					
English Background(EB)	11.58	3	3.86	.11	.05
Attitude (A)	4.98	2	2.49	.27	.02
EB x A	5.98	6	1.00	.79	.03
Errors	230.21	121	1.90		
Within Subjects					
Speech Samples (SS)	1369.40	5	273.88	.00	.78
SS x EB	12.79	15	.85	.18	.03
SS x A	5.49	10	.55	.57	.01
SS x EB x A	17.33	30	.58	.62	.04
Errors	386.31	605	.64		

TABLE 5
Summary of the Repeated Three-way ANOVA Using Analytic Scores for Grammatical Accuracy

Source	SS	df	MS	F	P
Between Subjects					
English Background(EB)	25.91	3	8.64	3.77*	.01
Attitude (A)	20.69	2	10.34	4.51*	.01
EB x A	8.96	6	1.49	0.65	.69
Errors	277.39	121	2.29		
Within Subjects					
Speech Samples (SS)	852.12	5	170.43	213.43**	.00
SS x EB	11.32	10	1.13	1.42	.17
SS x A	21.23	30	.71	0.89	.64
SS x EB x A	21.23	30	.71	0.89	.64
Errors	483.09	605	.80		

TABLE 6
Summary of the Repeated Three-way ANOVA Using Analytic Scores for Organization

Source	SS	df	MS	F	P
Between Subjects					
English Background(EB)	52.79	3	17.60	5.53**	.00
Attitude (A)	13.13	2	6.57	2.06	.13
EB x A	11.05	6	1.84	0.58	.75
Errors	385.24	121	3.18		
Within Subjects					
Speech Samples (SS)	682.17	5	136.43	164.20**	.00
SS x EB	25.43	15	1.70	2.04*	.01
SS x A	9.03	10	.90	1.09	.37
SS x EB x A	25.07	30	.84	1.01	.46
Errors	502.69	605	.83		

Post hoc analyses of the data for the analytic ratings on grammar revealed significant mean differences between NSs in Korea and NSs in the US, between NSs in Korea and NNSs in Hong Kong, between NSs in Korea and NNSs in Korea (see Table 7). The group of NSs in Korea, which consisted of more neutral and negative attitudes toward WEs in language testing than the positive, provided lower mean scores on grammar than other rater groups. Post hoc analyses of the rating data on organization also revealed significant mean differences between NSs in the US and NSs in Korea, between NSs in the US and NNSs in Hong, and between NSs in the US and NNSs in Korea (see Table 8). The group of NSs in the US, which consisted of more positive attitudes toward WEs in language testing than neutral or negative attitudes, provided higher mean scores on organization than other rater groups did. Even though raters' English language backgrounds have main effects on the analytic ratings on grammar and organization, the post hoc test results imply that raters' attitudes toward WEs play an underlying role in affecting rater's rating performance.

TABLE 7
Bonferroni Multiple Comparisons of Analytic Scores on Grammatical Accuracy

English language background	English language background	Mean Difference	Std. Error	Sig.
NS in the US	NS in Korea	.55*	.15	.00
	NNS in HK	.13	.16	1.00
	NNS in Korea	.14	.15	1.00
NS in Korea	NNS in HK	-.42*	.15	.04
	NNS in Korea	-.41*	.15	.03
NNS in HK	NNS in Korea	.02	.15	1.00

TABLE 8
Bonferroni Multiple Comparisons of Analytic Scores on Organization

English language background	English language background	Mean Difference	Std. Error	Sig.
NS in the US	NS in Korea	.77*	.18	.00
	NNS in HK	.78*	.19	.00
	NNS in Korea	.81*	.18	.00
NS in Korea	NNS in HK	.01	.18	1.00
	NNS in Korea	.04	.17	1.00
NNS in HK	NNS in Korea	.02	.18	1.00

In addition, the results of the repeated three-way ANOVA indicated that there were significant main effects for raters' attitudes toward WEs in language testing in analytic ratings on grammar, rate of speech, and task fulfillment, but not in holistic ratings (see Tables 4, 5, 9, & 10). Post hoc analyses of the data for the analytic ratings on grammar

revealed significant mean differences between positive and neutral raters and between positive and negative raters (see Table 11). The group of positive raters provided higher mean scores on grammar than other rater groups did. Post hoc analyses of the rating data on rate of speech also revealed significant mean differences between positive and negative raters and between positive and neutral raters (see Table 12). The group of positive raters provided higher mean scores on rate of speech than other rater groups did. Table 13 reveals that raters who had positive attitudes toward WEs in language testing provided the higher mean scores on task fulfillment than the negative and neutral rater groups. These results imply that neutral and negative raters have similar rating performance while positive raters have a significantly different rating performance in that the positive raters provide higher scores for certain analytic components such as grammar, rate of speech, and task fulfillment than negative or neutral raters do.

TABLE 9

Summary of the Repeated Three-way ANOVA Using Analytic Scores for Rate of Speech

Source	SS	df	MS	F	P
Between Subjects					
English Background(EB)	6.98	3	2.33	1.04	.38
Attitude (A)	15.22	2	7.61	3.39*	.04
EB x A	6.81	6	1.14	0.51	.80
Errors	271.60	121	2.25		
Within Subjects					
Speech Samples (SS)	921.03	5	184.21	227.99**	.00
SS x EB	43.39	15	2.89	3.58**	.00
SS x A	8.25	10	.83	1.02	.42
SS x EB x A	26.92	30	.90	1.11	.32
Errors	488.81	605	.81		

TABLE 10

Summary of the Repeated Three-way ANOVA Using Analytic Scores for Task Fulfillment

Source	**SS**	**df**	**MS**	**F**	**P**
Between Subjects					
English Background(EB)	20.42	3	6.81	1.98	.12
Attitude (A)	43.97	2	21.99	6.39**	.00
EB x A	13.86	6	2.31	0.67	.67
Errors	416.49	121	3.44		
Within Subjects					
Speech Samples (SS)	646.62	5	129.32	163.31**	.00
SS x EB	26.10	15	1.74	2.20*	.01
SS x A	21.53	10	2.15	2.72**	.00
SS x EB x A	20.19	30	.67	0.85	.70
Errors	479.08	605	.79		

TABLE 11
Bonferroni Multiple Comparisons of Analytic Scores on Grammatical Accuracy

Attitude	Attitude	Mean Difference	Std. Error	Sig.
Negative	Neutral	.14	.13	.88
	Positive	-.30	.13	.08
Neutral	Positive	-.44*	.13	.00

TABLE 12
Bonferroni Multiple Comparisons of Analytic Scores on Rate of Speech

Attitude	Attitude	Mean Difference	Std. Error	Sig.
Negative	Neutral	.14	.13	.91
	Positive	-.48*	.13	.00
Neutral	Positive	-.34*	.13	.03

TABLE 13
Bonferroni Multiple Comparisons of Analytic Scores on Task Fulfillment

Attitude	Attitude	Mean Difference	Std. Error	Sig.
Negative	Neutral	.05	.16	1.00
	Positive	-.59*	.16	.00
Neutral	Positive	-.64*	.16	.00

V. DISCUSSION

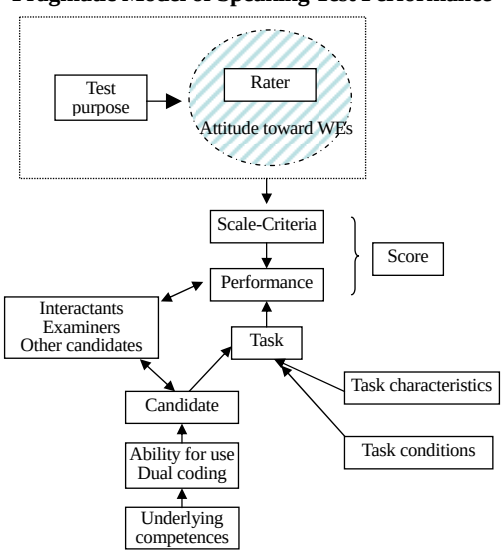
As L2 oral tests are increasingly used in ESL/EFL contexts, the validation process should be laid down as comprehensively as possible in order to ensure test scores are used or interpreted appropriately. The investigation of rater variability in the assessment process of L2 oral tests helped to achieve validity of test score interpretation/use in that the sources of invalidity were found in a specific way of raters' rating process.

The present findings reveal how raters' attitudes toward WEs in language testing influence their rating performances. Raters' different attitudes toward WEs significantly affected the analytic ratings on grammar, rate of speech, and task fulfillment. Findings in the present study could be considered more informative for understanding the rater effect on the ratings of L2 oral tests than the previous research (Brown, 1995, 2004; Chalhoub-Deville, 1995a, 1995b; Galloway, 1980; Lumley & McNamara, 1993; Smith & Bisazza, 1982) since not only raters' English language backgrounds, but also their attitudes toward WEs in language testing have been reported as important factors influencing test scores. The present study reveals that NSs in the US hold more positive attitudes toward WEs in language testing compared to the other rater groups (NSs in Korea, NNSs in Hong Kong, and NNSs in Korea) and show leniency on the ratings of Korean students' speech

samples. Attitudes toward WEs as well as raters' English language backgrounds were important factors that influence some analytic test scores. Therefore, in terms of attitudes toward WEs, it is less important to guard against whether raters are native speakers of English or not. Given the present findings, where the type of attitude raters hold with regard to WEs seemed to influence some of their analytic ratings, test developers and researchers should monitor and document raters' attitudes toward WEs to help ensure more appropriate scores.

The problems of using a single standard English for the assessment of non-native speakers' English for international communication are presented in this study as argued in the WEs field (Brutt, 2001; Brutt-Griffer & Sammy, 1999; Davies, 1990; Pennycook, 2001; Widdowson, 1994). It is necessary to develop an appropriate assessment framework according to rater groups in order to enhance the validity of test score interpretation and use. Therefore, researchers in the language testing field should consider in detail the potential rater characteristics that influence test scores and conduct further research on the effects of these raters' characteristics. These characteristics may include the impact of raters' attitudes toward WEs on their perceptions of rating criteria for the evaluation of non-native speakers' English language oral proficiency and on the assessment process with different test-tasks. In terms of WEs, it is important to point out that depending on test purpose, it may not be appropriate to use a single standard English for the assessment of non-native speakers' English language oral proficiency. I suggest a slightly modified model for the assessment process for non-native speakers' English language oral proficiency based on Skehan's (1998) model as shown in Figure 3.

FIGURE 3
Pragmatic Model of Speaking Test Performance



In sum, raters cannot be grouped by their English language backgrounds and generalized as homogeneous raters. Since there were significant differences among individual raters within the same rater groups categorized by their English language backgrounds, this study suggests that the ratings of L2 oral tests should identify "who the raters are," "which characteristics they have," and "which rating components were salient in their rating process" in order to provide appropriate test score interpretation. Especially, the present findings reveal that raters' attitudes toward WEs are related to their rating scheme. Therefore, research on the rater effect on test scores within the WEs perspective is important to suggest a new assessment framework and to enhance validity of test score interpretation and use. Further research on rater effects on test scores with various raters residing in different ENL, ESL, and EFL countries is needed.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Brumfit, C. J. (2001). *Individual freedom in language teaching: Helping learners to develop a dialect of their own.* Oxford: Oxford University Press.

Brutt-Griffler, J., & Samimy, K. K. (1999). Revisiting the colonial in the postcolonial critical praxis for nonnative English-speaking teachers in a TESOL program. *TESOL Quarterly, 33*, 413-431.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*, 369-383.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Researching pedagogic tasks: Second language learning, teaching and testing.* Harlow, English: Pearson Education Limited.

Davies, A. (1990). *Principles of language testing.* Cambridge, MA: Basil Blackwell.

Jenkins, J. (2003). *World Englishes*. London and New York: Routledge

Kachru, B. B. (1982). *The other tongue*. Urbana: University of Illinois Press.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in outer circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the World: Teaching and learning the language and literatures* (pp. 11-30). Cambridge: Cambridge University Press.

Kachru, B. B. (1994). Englishization and contact linguistics, *World Englishes, 13*(2), 135-154.

Kachru, B. B. (1997). Past imperfect: The other side of English in Asia. In L. Smith & F. Michael (Eds.), *World Englishes 2000: Resources for research and teaching* (pp. 209-251). Hawaii: University of Hawaii.

Kachru, B. B., & Nelson, C. L. (2001). World Englishes. In A. Burns & C. Coffin (Eds.), *Analyzing English in a global context* (pp. 9-25). New York: Routledge.

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.

McArthur, A. (1987). The English languages? *English Today, 11*, 9-13.

McKay, S. (2002). *Teaching English as an international language: Rethinking goals and approaches.* Oxford: Oxford University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London: Addison-Wesley Longman.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446-466.

McNamara, T. F. (2003). Looking back, looking forward: Rethinking Bachman. *Language Testing, 20*, 466-473.

Modiano, M. (1999a). International English in the global village. *English Today, 15*(2), 22-34.

Modiano, M. (1999b). Standard English (es) and educational practices for the world's lingua franca. *English Today, 15*(4), 3-13.

Pennycook, A. (1999). Pedagogical implications of different frameworks for understanding the global spread of English, In C. Gnutzmann (Ed.), *Teaching and learning English as a global language.* Tubingen: Stauffenburg Verlag.

Pennycook, A. (2001). *Critical applied linguistics: A critical introduction*. Mahwah, NJ: Erlbaum.

Seidlhofer, B. (1999). Double standards: Teacher education in the expanding circle. *World Englishes, 18*(2), 233-245.

Shim, R. J. (1999). Codified Korean English: Process, characteristics and consequence. *World Englishes, 18*(2), *247-258.*

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Thorndike, E. L. (1920). A Constant error on psychological rating. *Journal of Applied Psychology, 6*, 25-29.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing,*

11, 197-223.

Widdowson, H. G. (1994). The ownership of English. *TESOL Quarterly, 28*(2), 377-389.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305-335.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks second language learning, teaching and testing* (pp. 186-209). Essex, UK: Longman.

Yano, Y. (2001). World Englishes in 2000 and beyond. *World Englishes, 20*(2), 119-131.

Applicable levels: college and university
Key words: World Englishes (WEs), English language oral proficiency, assessment process, rating scales, attitudes

Hyun-Ju Kim
217 Hanyang Integrate Technology (HIT)
17 Hengdang-dong, Sungdong-gu
Hanyang Cyber University
Seoul city, S. Korea.
Tel: (02) 2290-2944
E-mail: hyunjuk@gmail.com

Received in August, 2005
Reviewed in September, 2005
Revised version received in November, 2005