Exploring the assessment of English pronunciation using GENOVA and FACETS

Hiroko Yoshida
Osaka University of Economics

Abstract

This study investigates the potential for the analytic assessment of the English pronunciation of Japanese using GENOVA (Crick & Brennan, 1984) and FACETS (Linacre, 1996a). A total of 21 Japanese EFL college students read two different materials (a prose type reading and a dialog type reading) and these were audiotaped. Their pronunciation performances were rated by five judges (three L1 Japanese and two L1 English instructors) using 15 assessment items (vowels, diphthongs, consonants, consonant clusters, aspiration, word stress, sentence stress, rhythm, intonation, weak forms, loudness, tempo, energy, smoothness, and clarity). The results revealed that the performance-based analytic pronunciation assessment served its purposes even though the raters made their judgments independently. The study also showed that (1) The 15 items significantly varied in difficulty and (2) The raters exerted different levels of severity in rating.

Key Words: pronunciation, assessment, GENOVA, FACETS

1. Introduction

Pronunciation has been placed in a prominent position in the ESL/EFL classroom after it experienced a number of ups and downs in the language curriculum (Celce-Murcia, Brinton, & Goodwin, 1996; Morley, 1991). There are three causes for the increased awareness of pronunciation. Firstly, with the advent of communicative competence, the new perspectives on language learning that encompass wider aspects of language have been supported by many TESOL professionals. Consequently, pronunciation has been recognized as having an important role in communication. The development of research on discourse analysis has also contributed to the new role of pronunciation in receptive and productive communication. For example,

suprasegmentals, such as stress, rhythm, and intonation and paralinguistic features, such as loudness and articulation can convey the speakers' intentions and emotions (Brown, 1990). Moreover, the role of pronunciation that serves as a navigation guide for the listener to follow communication (Gilbert, 1994) has been recognized. Pennington and Richards (1986) clearly explained this newly recognized role of pronunciation in communication; "(p)ronunciation is seen not only as part of the system for expressing referential meaning, but also as an important part of the interaction dynamics of the communication process" (p. 208), and "Pronunciation is not simply a surface performance phenomenon but is rather a dynamic component of conversation fluency" (p. 212). Furthermore, the importance of pronunciation has also been urged outside the language classroom. Along with an increasing number of non-native speakers of English who speak English with native/non-native speakers, various types of communication breakdowns or misunderstandings caused by marked accents have been reported (Anderson-Hsieh & Koehler, 1988; Eisenstein, 1983; Fayer & Krasinski, 1987; Madden & Moore, 1997; McKenna, 1987; Munro & Derwing, 1995). As a result, many language professionals have acknowledged the need for teaching pronunciation.

However, a growing interest in pronunciation teaching has primarily focused on "content teaching strategies and materials" (Goodwin, Brinton, & Celce-Murcia, 1994). Whilst a few works addressed pedagogical assessments to examine pronunciation for diagnostic purposes (Celce-Murcia et al., 1996; Goodwin et al., 1994; Morley, 1988), the systematic study of the performance tests on pronunciation has been extremely limited. Therefore, the purpose of this study is to explore the potential for examining pronunciation performance by using two approaches, which have been discussed in the body of literature in the field of language testing: generalizability theory (Bachman, Lynch, & Mason, 1995; Bolus, Hinofotis, & Bailey, 1982; Brennan, 1983, 2001; Brown, 1999; Brown & Ross, 1996; Shavelson & Webb, 1991) and multifaceted Rash analysis (Kondo-Brown, 2002; Linacre, 1996b; Lumley & McNamara, 1995; McNamara, 1996; Weigle, 1998). The present study addresses the following three questions.

1. To what degree and in what way do the facets of the analytic pronunciation performance test (tasks, raters, assessment items) contribute to scores?

2. To what degree does the rating scale function sufficiently in evaluating pronunciation performances?
3. To what degree is the pronunciation evaluation reliable?

2. Method

2.1. Participants

Twenty one second-year female students at a junior college in Osaka participated in this study. They were enrolled in the researcher's intermediate-level English class where they learned recent international news in English. Their ages ranged from 18 to 20 years old and they were homogenous with regard to educational background (12-13 years of formal education in Japan). All students were L1 Japanese speakers, majoring in English.

2.2. Materials and evaluation items

Two materials were used in this study: a dialog type reading (D task) and a prose type reading (P task)[1]. The readings are shown in the Appendix. These materials were taken from *Accurate English* (Dauer, 1993). Three aspects of the sound system of General American English (GAE), segmentals, suprasegmentals, and paralinguistic features were examined. The segmental aspect of language refers to individual sounds, and particular combinations of the individual sounds. Suprasegmentals are the features beyond one sound segment (Celce-Murcia et al., 1996). The paralinguistic features refer to features that are "considered to be beyond the set of phonological contrasts of a language" (Roach, 2001). Items in each category are: vowels, diphthongs, consonants, consonant clusters, and aspiration in the category of segmentals; word stress, sentence stress, rhythm, intonation, and weak forms in the suprasegmentals category; and loudness, tempo, energy, smoothness, and clarity in the paralinguistic features (For detailed information for these items, see Yoshida, 2005). The 15 items were selected because (a) these items are difficult for Japanese learners to acquire (Avery & Ehrlich, 1992; Dale & Poms, 1994; Thompson, 1987), and (b) these items are crucial aspects of "phonological intelligibility" (Jenkins, 2000, p. 123).

2.3. Procedure

The participants audiotaped their readings of the two tasks in the language laboratory (Sony, ER-9030) during the regular class time at the college. Each student

used a headset with a microphone (Sony, HS-90). The researcher dubbed the recorded tapes on two tapes randomly, so that one tape contained all students' performances of the dialog readings that were randomly ordered, and the other included those of the prose readings randomly ordered. Two kinds of tape were handed to three L1 Japanese and two L1 English raters, respectively.

All L1 Japanese raters were fluent English speakers who have Master of Education degrees from a US. University (Raters 1, 2, and 3). The researcher participated in the study as Rater 3. Two L1 English raters were graduate students in the Master of Education Program (Raters 4 and 5). Both L1 Japanese and English raters had completed a phonology class; therefore they were familiar with the basic sound system of GAE. Before the ratings, individual rater guidance was provided where raters practiced sample ratings[2]. Raters were asked to judge a dialog reading performance first, and a prose reading performance in this order[3], respectively, using a 6-point Likert scale; 1 = *Very poor*, 2 = *Poor*, 3 = *Fair*, 4= *Good*, 5 = *Very good*, and 6 = *Excellent*. Raters evaluated the two reading tasks according to the 15 phonological elements of GAE. The ratings were conducted at each rater's home after the rater guidance was completed. After finishing the ratings, the raters were asked to provide qualitative feedback on their ratings.

3. Results

3.1. Overview of the analysis

Multifaceted Rash analysis was used to analyze the data, using the computer program FACETS (Linacre, 1996a). Figure 1 presents graphically the measures for person ability, task difficulty, rater severity, and item difficulty. The scale in the leftmost column represents the logit scale. In all facets in Figure 1, the same logit scale is used to illustrate a continuum of ability, severity, and difficulty by the analysis. Each person is identified by her ID number (1-21) in the second column. Persons shown at the higher logit mean they are more able. For, example, Person 21 was the most able among all participants. The third column shows the task difficulty. The higher the rating on the scale, the more difficult. The fourth column shows rater severity. Raters are ordered with the most severe at the top, and the least severe at the bottom. The fifth column presents the difficulty variation among items. Similarly, the most difficult item was uppermost, and the least difficult item was at the bottom.

As Figure 1 shows, person ability estimates are relatively clustered (from a high of about 1 logit to a low of close to -1 logit). The facet of task shows that the prose reading task was slightly more difficult than the dialog. For raters, there was variation in severity ranging approximately from -1 to +1 on the logit scale. Eight items with negative logit scales showed that they were easier than average, and seven items with positive logit scales showed that they were more difficult than average.

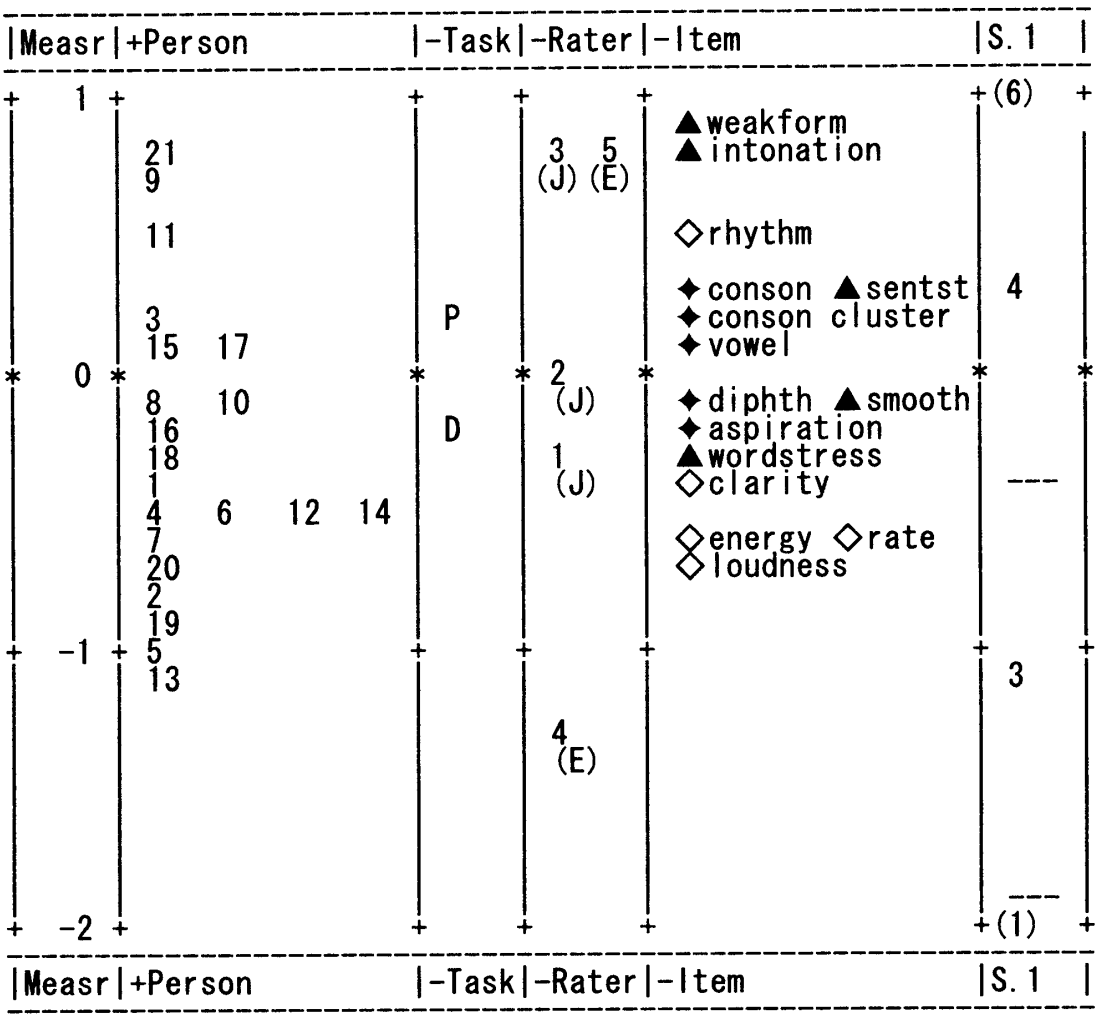

Figure 1
FACETS Summary
Notes. ✦ segmentals, ▲suprasegmentals, ◇ paralinguistic features, ✦conson: consonant, ▲sentst: sentence stress, ✦conson cluster: consonant cluster, ✦diphth: diphthong, ▲smooth: smoothness

3.2. Task difficulty

Table 1 provides estimates of task difficulty. Each column presents task, task

difficulty, error, and infit mean square value, from left to right. The second left column shows that the dialog task had -0.24 logit, and the prose task had 0.24 logit. Although the absolute difference between two logit values was small (.48), the reliability index, which presents the extent to which the test distinguishes the difficulty of tasks, was high (.98). Furthermore, the chi-square, which examines the null hypothesis that all tasks were equal, indicated 108.80 with *df* 1, which was significant at $p = .00$. Thus, these findings suggest that the two tasks varied in difficulty.

Table 1

Difficulty Measurement Report for Two Tasks

Task	Difficulty (logits)	Error	Infit (mean square)
Dialog reading	-0.24	0.03	1.10
Prose reading	0.24	0.03	0.90

3.3. Rater severity and consistency

Table 2 provides a more detailed analysis of rater behavior. Each column presents rater, rater severity, error, and infit mean square values, from left to right. Raters are presented in descending order of leniency. Rater 4 (L1 English rater) was the most lenient, and Rater 3 (L1 Japanese rater) was the most severe. The reliability index was very high (1.00), and the chi-square of 1110.70 with *df* 4 was significant at $p = .00$. Therefore, the hypothesis that the raters were equally severe was rejected. Put differently, the five raters showed a large range of severity in their rating, with Rater 4 the most lenient (-1.32), and Rater 3 the most severe (0.83).

In Table 2 the infit statistic shows whether or not each rater was consistent with their ratings. McNamara (1996, p. 173) reports that the lower and upper limits of .75 and 1.30, respectively, is acceptable. In Table 2, Rater 5 had a very high infit statistic (1.50), showing that Rater 5 was not consistent with her rating, thus, the scores that she gave lacked predictability.

3.4. Item difficulty

Table 3 provides estimates of item difficulty. The items are presented in

descending order of ease. The item of loudness was the easiest (-.70), and the weak form item was the most difficult (.88). The reliability index was very high (1.00), and the chi-square of 411.20, with *df* 14 was significant at p= .00. Thus, these findings suggest that the 15 items varied in difficulty.

Table 2

Rater Measurement Report for Five Raters

Rater	Severity (logits)	Error	Infit (mean square)
4(E)	-1.32	0.06	0.90
1(J)	-0.34	0.05	0.80
2(J)	0.01	0.05	0.80
5(E)	0.82	0.05	1.50
3(J)	0.83	0.05	0.90

Note. (E): L1 English, (J):L1 Japanese

Table 3

Difficulty Measurement Report for 15 Items

Item	Difficulty (logits)	Error	Infit (mean square)
loudness	-0.70	0.09	0.90
rate	-0.58	0.09	0.80
energy	-0.58	0.09	1.00
clarity	-0.39	0.09	1.00
word stress	-0.33	0.09	0.90
aspiration	-0.20	0.09	0.80
diphthong	-0.09	0.09	0.80
smoothness	-0.09	0.09	1.00
vowel	0.14	0.09	0.90
consonant cluster	0.19	0.09	1.10
consonant	0.25	0.09	1.10
sentence stress	0.28	0.09	1.20
rhythm	0.46	0.09	1.00
intonation	0.76	0.09	1.20
weak form	0.88	0.09	1.30

3.5. Rating scale

Table 4 shows how frequently each rating scale was used in rating. The first column presents the rating scale. The second and third columns show the number and the percentage of observed use of each category, respectively. The fourth column presents step difficulty, which indicates the log-ratio of the frequency of adjacent categories. The rightmost column indicates the logit measure for the expected score corresponding to the value in the category score column. The step difficulty advances from scale 5 to 6 by 5.95 logits, suggesting the interval between scale 5 and 6 was too wide, and scale 6 was less informative as a scale (Linacre, 1997).

Table 4

Frequency Measurement Report for Rating Scale

Rating scale	Count	%	Step difficulty	Category measure
1	49	2		-4.91
2	455	14	-3.75	-2.83
3	952	30	-1.74	-1.14
4	1050	33	-0.46	0.33
5	631	20	0.90	2.99
6	13	0	5.05	6.13

Note. 1: *Very poor* 2: *Poor* 3: *Fair* 4: *Good* 5: *Very good* and 6: *Excellent*

3.6. Generalizability coefficients (reliability)

Generalizability theory (G-theory) was used in analyzing the reliability of this study. All analyses were performed using the GENOVA program (Crick & Brennan, 1984). G-theory extends the notion of reliability that is calculated in traditional classical theory. In classical theory, an observed measurement consists of two elements: a "true" score and a single undifferentiated "error", whereas G-theory allows us to sort out multiple sources of error and to provide a more comprehensive explanation of the relative importance of various sources of error (Brennan, 2001, p. 2-3). Consequently, G-theory makes it possible to decide which measurement conditions will be relevant on test performance data (Lynch & McNamara, 1998). The estimated reliability coefficients called G coefficients can be obtained for both norm-reference (NRT) and criterion-reference (CRT) interpretations. Furthermore,

the estimated G coefficients can be analyzed to determine the optimal numbers of rating conditions, such as the number of raters. The estimated reliability of the present study for the dialog reading task for NRT based on five raters and 15 items was .72760, and for CRT, .54463. The estimated reliability for the prose reading task was .44463 for NRT, and .22972 for CRT.

As the estimated G coefficients of the prose reading task for this study were relatively low, the raters were divided into two groups according to their L1: L1 English and L1 Japanese groups. Whether the G coefficients were similarly low for both groups was investigated. For comparison, the G coefficients of the dialog reading task were examined for the two groups. The results are visually shown in Figure 2 and Figure 3. As these figures clearly show, the L1 Japanese raters showed a higher reliability than the L1 English raters in both the dialog and prose reading tasks for NRT and CRT purposes. For NRT, three L1 Japanese raters had a .84 G coefficient in the dialog reading task, while two L1 English raters obtained only .14. In figure 3, the difference between the L1 Japanese raters and L1 English raters are more clearly presented. While the reliability increases when the number of L1 Japanese raters increases for both NRT and CRT, the reliability of the L1 English raters was constantly low when the number of raters increases for both tasks.

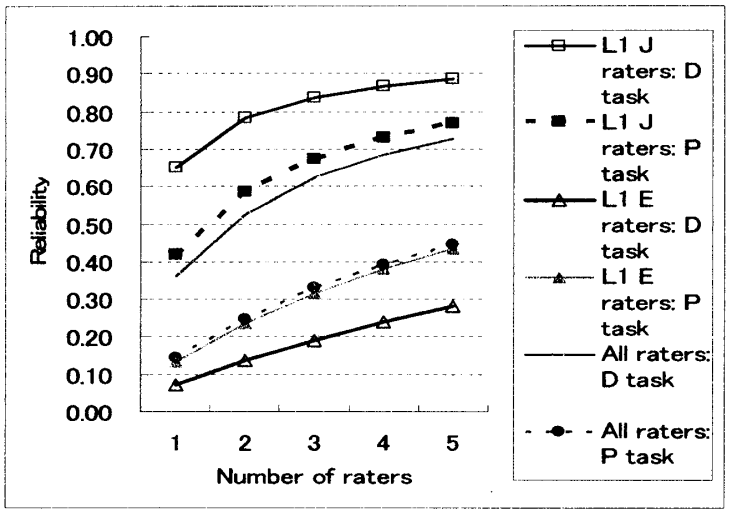

Figure 2
G Coefficients of Two Rater Groups for NRT

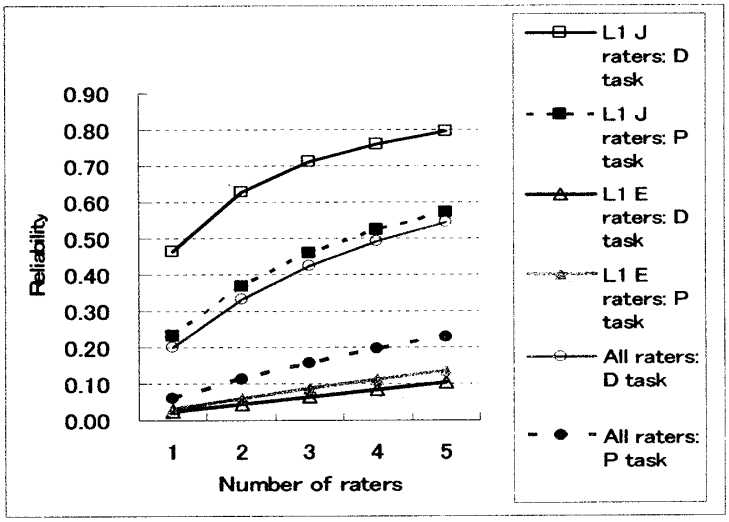

Figure 3
G Coefficients of Two Rater Groups for CRT

4. Discussion

This study revealed the potential for evaluating pronunciation performance. First of all, although the dialog reading task was easier for the participants, it elicited more reliable judgments compared with the prose reading task. Second, despite uniform rater guidance provided for every rater, the raters showed different levels of severity in rating. Qualitative feedback obtained after this study revealed that the L1 Japanese raters tended to strictly follow rating references, while the L1 English raters had relative difficulty in the avoidance of making judgments based on their L1 pronunciation experience. However, the varying degrees of severity were attributed to individual rater differences rather than L1 background. The results suggest that more comprehensive rater guidance taking into account different rater experience, beliefs, and background may be necessary in the future research examining pronunciation performance.

In terms of item difficulty, the analysis shows varying levels of difficulty among items. However, the difficulty was not ordered according to the three different elements of GAE: segmentals, suprasegmentals, and paralinguistic features. One argument proposed by raters was the use of microphone when recording pronunciation performances. As the microphone was rather sophisticated, it recorded all voices clearly, yet some performances might have lacked clarity, energy or

loudness in a real communication context. Therefore, paralinguistic features may tend to be overevaluated in this rating situation. The critical role of paralinguistic features is discussed (Pennington & Richards, 1986); however, the results of this study suggest the difficulty in precisely evaluating it in the language laboratory.

Furthermore, the analysis shows that a 6-point Likert scale did not sufficiently function as expected because the scale 6 was less informative as a scale. Qualitative feedback from raters supported this result: three Japanese raters noted that it was difficult to distinguish "Excellent" from "Very good" in English.

Finally, the results show the assessment was relatively reliable in the dialog reading task with the present 5 raters. In addition, the increase in the number of raters can boost the reliability of the pronunciation assessment.

5. Conclusion

This study examined a pronunciation performance assessment. The performance-based analytic pronunciation assessment served its purposes, even though the raters made their judgments independently. However, future tasks to implement more reliable and dependable pronunciation assessment have been revealed. This study reveals that each facet of tasks, raters, assessment items of the assessment can have varying impact on judgments; therefore, it is crucial that test developers should carefully take these factors into consideration, and pay special attention to minimize those effects especially in rater guidance. Finally, it is necessary to keep in mind the limitations of this study. Because the number of participants of the present study was small and all of them were selected from streamed intact classes, further studies with larger participants consisting of various proficiency levels are needed to substantiate the results. It is hoped that the findings of this study will be incorporated into the full-scale development of an analytic instrument for evaluating pronunciation.

References

Anderson-Hsieh, J., & Koehler, K. (1988). The effects of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*, 561-603.

Avery, P., & Ehrlich, S. (1992). *Teaching American English pronunciation.* 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*, 238-257.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa, IA: American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brown, G. (1990). *Listening to spoken English*. London: Longman.

Brown, J. D. (1999). The relative importance of persons, items, subtests, and languages to TOEFL test variance. *Language Testing, 16*, 217-238.

Brown, J. D., & Ross, J. A. (1996). Decision dependability of subtests, tests and the overall TOEFL test battery, *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem*. Cambridge: Cambridge University Press.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge: Cambridge University Press.

Crick, J. E., & Brennan, R. L. (1984). A general purpose analysis of variance system (Version 2.2) [Computer software]. Iowa City, IA: American College Testing Program.

Dale, P. W., & Poms, L. (1994). *English pronunciation for Japanese speakers*. Englewood Cliffs, NJ: Prentice Hall.

Dauer, R. M. (1993). *Accurate English: A complete course in pronunciation*. Englewood Cliffs, NJ: Prentice Hall.

Eisenstein, M. (1983). Native reactions to non-native speech: A review of empirical research. *Studies in Second Language Acquisition, 5*, 160-176.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313-326.

Gilbert, J. B. (1994). Intonation: A navigation guide for the listener. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 36-48). Bloomington, IL: TESOL.

Goodwin, J., Brinton, D., & Celce-Murcia, M. (1994). Pronunciation assessment in the ESL/EFL curriculum. In J. Morley (Ed.), *Pronunciation pedagogy and theory:*

New views, new directions (pp. 3-16). Washington D. C.: TESOL.

Jenkins, J. (2000). *The phonology of English as an international language.* Oxford: Oxford University Press.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3-31.

Linacre, J. M. (1996a). Facets (Version 3.0) [Computer software]. Chicago: MESA.

Linacre, J. M. (1996b). *A user's guide to Facets.* Chicago: MESA Press.

Linacre, J. M. (1997). *MESA Note 2: Guidelines for rating scales.* Retrieved October, 17, 2002, from http://209.41.24.153/rn2.htm

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

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*, 158-180.

Madden, M., & Moore, Z. (1997). ESL Students' opinions about instruction in pronunciation. *Texas Papers in Foreign Language Education, 3*, 15-32.

McKenna, E. (1987). Preparing foreign students to enter discourse communities in the U.S. *English for Specific Purposes, 6*, 187-202.

McNamara, T. F. (1996). *Measuring second language performance.* New York: Longman.

Morley, J. (1988). How many languages do you speak? Perspectives on pronunciation-speech-communication in EFL/ESL. *Nagoya Gakuin Daigaku Gaikokugo Kyoiku Kiyo, 19*, 1-33.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly, 25*, 481-520.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*, 73-97.

Pennington, M. C., & Richards, J. C. (1986). Pronunciation revised. *TESOL Quarterly, 20*, 207-225.

Roach, P. (2001). *Phonetics.* Oxford: Oxford University Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Thompson, I. (1987). Japanese speakers. In M. Swain & B. Smith (Eds.), *Learner English: A teacher guide to interference and other problems* (pp. 212-223). Cambridge: Cambridge University Press.

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

Yoshida, H. (2005). Validity of an instrument measuring English pronunciation performance. *JACET Kansai Journal, 8*, 1-14.

Appendix

A dialog type reading task

Ms. Green: Come in, Taro. How are you?
Taro: Fine, Ms. Green. I'd like to ask you an important question. You'll answer it honestly, won't you?
Ms. Green: Well, I'll try.
Taro: How's my accent?
Ms. Green: What do you think?
Taro: I've studied hard in your course, but I don't know.
Ms. Green: I think you've made a lot of improvement.
Taro: Improvement? But do I sound like an American?
Ms. Green: Not exactly. But you don't have to, because you are Japanese.
(Dauer, 1993, p. 241, Adapted with permission of Prentice-Hall/Addison-Wesley Longman)

A prose type reading task

Learning to speak a foreign language fluently and without an accent isn't easy. In most educational systems, students spend many years studying grammatical rules, but they don't get much of a chance to speak. Arriving in a new country can be a frustrating experience. Although they may be able to read and write very well, they often find that they can't understand what people say to them. English is especially difficult because the pronunciation of words is not clearly shown by how they are written. But the major problem is being able to listen, think, and respond in another language at a natural speed. This takes time and practice.

(Dauer, 1993, p.6)

---

[1] For a dialog reading, the researcher reduced the length of the passage from 93 words to 70 words and modified the passage.
[2] One rater, Rater 5, was asked to practice sample rating at home after rater guidance because of time constrains.
[3] It is ideal to use counterbalancing in rating procedure; however, it is difficult to rigorously control the order of ratings because all ratings were individually conducted at home in this study. Therefore, counterbalancing was not used in ratings.