Backwash of Performance Testing: Auditory and Acoustic Analyses of an Utterance Recorded Before and After an EFL Pronunciation Course

SHIZUKA, Tetsuhito
Kansai University

Abstract

The present paper reports on auditory and acoustic differences observed in learners' utterances before and after a series of pronunciation tests. A group of Japanese junior high school students (*N* = 66) participated in a 24-session pronunciation course, in which the main activity was one-on-one performance coaching/testing. Audio recordings were made of a short sentence read aloud by a subgroup (*n* = 34) of the participants both at the beginning of Session 1 and at the end of Session 24. The pre- and post-course recordings were compared in terms of (a) perceived degree of holistic foreign accentedness, (b) perceived phonological accuracy of segments, (c) the range and *SD* of vocal pitch (F0), and (d) acoustic characteristics of /r/ as reflected in F3. Holistic accentedness was judged by L1 English speakers, segments were rated by trained L1 Japanese teachers of English, and F0 and F3 analyses were conducted using Praat speech analysis software. The results indicated that in terms of every variable measured the post-course recordings were closer to the targeted model than the pre-course recordings were. Through the one-year training, participants' utterances were segmentally more accurate, lower in the degree of foreign accent, came to use a wider pitch band, and F3 frequencies for /r/s became lower. Significance of these results is discussed and an argument is presented that the observed effects should be considered beneficial backwash of performance testing of pronunciation.

Key words: pronunciation, one-on-one performance testing, auditory and acoustic analyses

1. Introduction

Anecdotes abound about communication breakdown and awkward moments attributable to words pronounced poorly by Japanese learners of English. Nonaka (2005a) cites a Japanese college student who was given a can of Budweiser by her host-father after she commented on the "bad weather" of the day. Jenkins' (2000) nonnative-nonnative dyad interaction data include instances of non- or misunderstanding caused by Japanese versions of such simple words as "grey," "wood," "hat," "sad," and "Japan" (pp. 85-86).

It is scarcely difficult to come across Japanese learners of English who are strongly influenced by their native language at segmental, supra-segmental (Flege, 1980; 1981) and voice quality setting (Esling & Wong, 1983) levels. An early study by Suter (1976) found Japanese speakers of English much less intelligible than their Arabic, Persian, and Thai counterparts. Typically, they are poor at crucial phoneme distinctions (Jenkins, 2000; Arimoto, 2005), particularly at differentiating between /r/ and /l/ (Flege, Takagi, & Mann, 1995; Riney, Takagi, & Inutsuka, 2005), as well as at using a wide-enough pitch range (Todaka, as cited in Celce-Murcia, Brinton, & Goodwin, 1996). Nakanishi (2004) reports that Japanese students themselves find "Japanese English" less comprehensible as well as less attractive than native speaker varieties.

One might hope that the advent of World Englishes concept (Smith, 1983) is easing Japanese learners', as well as their teachers', predicament. With non-native speakers of English in the outer and the expanding circles (Kachru, 1985) outnumbering native speakers by the ratio of approximately 3 to 1 (Crystal, 2003), those who are teaching English to Japanese students might be tempted to convince themselves that the uphill, or losing, battle against "Japanese English" need no longer be wag That conclusion, however, seems largely misguided. Jenkins (2000) has proposed the Lingua Franca Core (LFC) as a more practical and suitable goal for most learners of English. A close examination of the LFC reveals that virtually all the phonological features of the English language at which Japanese learners are known to be weak are retained in the list. That is, it turns out that all those sounds Japanese learners have to struggle with are crucial even in communication with other groups of non-native speakers. In fact, phonetic deviations seem to cause more of a problem for non-native listeners than for native counterparts because the former have a narrower band of phonetic tolerance (Jenkins, 2000, p. 37). If that is so, exactly because of the expected proliferation of English-medium communication opportunities that do not involve any native speaker parties, skills to satisfactorily articulate at least core phonological items have become more important than ever.

Unlike dance steps, sounds of a new language cannot be learned simply by watching the instructor's movements: unlike a dancer's arms and legs, a language teacher's articulatory organs are mostly not in view (Yamada, Adachi, & ATR Institute, 1999). For that reason, as Arimoto (2005) rightly points out, the teacher's role as a coach is critical in pronunciation teaching. However, empirical studies on the effects of teacher intervention on improving Japanese students' pronunciation are rare. One of the few reports that the author has come across, by Asami and Tanaka (2005), is rather sketchy, making it difficult to interpret their results. Researchers and practitioners concur (Makino, 2005; Asami & Tanaka, 2005; Kosuge, 2005) that a major challenge in pronunciation teaching lies in maintaining learners' motivation until their knowledge is proceduralized, rather than in imparting declarative knowledge of articulatory phonetics. Unfortunately, proposals of systematic ways for enhancing and maintaining students motivation towards pronunciation are again hard to find.

2. One-on-One Testing as a Motivator

Possibly, the only known attempt in that direction is the personal card method (PCM) reported by Shizuka (1995). In the PCM, each of the students receives a personal card that specifies an appropriate number of pronunciation items--phrases or short sentences including target sounds—and they are told to train themselves outside class to master them. The students are encouraged to approach the teacher whenever and wherever possible out of class and to try orally producing any of the items, before the due date, which is, say, a week away. When a student's performance on an item is judged to be satisfactory, the student earns a point for that item. The total points earned before the due date will be the student's pronunciation score, which will account for a certain percentage in the final term grade. Hence, the essence of the PCM is cyclical one-on-one procedural-knowledge testing, as opposed to one-shot collective declarative-knowledge testing. Shizuka (1995) found that using this method three times over a nine-week period was effective in significantly improving the accuracy rate of college students' /r/ sounds from 44 to 80 percent, as well as in enhancing their self-reported motivation towards bettering pronunciation.

The current study constitutes the second part of an investigation into the effects of what could be termed a more intensive version of Shizuka's PCM. In the academic year 2007, the author taught a 24-session course in English pronunciation for junior high school 3rd-year students. The aim of the program was to improve participants' skills in key segmental and supra-segmental features of English phonology. Two approaches were taken to measure the effects of the course. One was continually administering a Likert scale survey to track possible changes in participants' self-perceived abilities and motivations. The other was audio-recording students' pronunciation of the same formulaic expression before and after the 24-week course, to submit the recordings to auditory and acoustic analyses. The survey results (Shizuka, in press) indicated that participants' self-perceived pronunciation ability gradually increased throughout, resulting in much higher perceived skills at the end of the course than at the beginning. It was also found that the course was instrumental in strengthening participants' motivation toward acquiring better pronunciation as well as heightening their pronunciation-attentiveness even outside the course. The purpose of the present paper is to complement the self-perception study by reporting on the results of auditory and acoustic analyses.

3. The Study

3.1 Participants

Participants of the pronunciation course were 66 (31 males and 35 females) 3rd-year students (age 14-15) at a private junior high school in western Japan. These students signed up for the elective course on a forced-choice basis; all the students in the 3rd year were required to choose one of the five elective courses offered in different subject domains. The junior high school was attached to a high school, which in turn was attached to a university. This ensured that the students were generally free from the common pressure to cram for entrance examinations to high schools. The 66 students were divided into two groups (32 and 34) to participate in the course consecutively on the same days. The first group were trained in the first two 50-minute periods and the second group in the second two 50-minute periods on the same mornings. The course contents for the two groups were identical. Only the second group (n = 34) made the recordings (see below) to be analyzed in this study.

3.2 Course Details

The course consisted of 24 sessions, 10 in Term 1, 9 in Term 2, and 5 in Term 3. One session lasted 100 minutes, with a 10-minute break in the middle. Course materials, which were 21 Handouts and 21 Personal Cards, were written by the author. A Handout presented short utterances from a movie clip and/or selected lines from a pop song that contained target phonological features of the day. A Personal Card listed seven or eight "items," which were utterances or song lines selected from the Handout. Main segmental targets were consonants known to pose difficulties to Japanese learners (e.g., /r/, /f/, /v/, /θ /, /ð/), though a few vowels (/æ/ /ə:/) were focused on as well. The first 21 sessions had only segmental features as explicit testing points, treating supra-segmental features like word stress, linking, intonation only incidentally, while the last three included stress-timed rhythm among focal points.

A typical structure of one session was as follows: the first 15 to 20 minutes were spent on teacher-fronted explanations and after-the-model collective repetitions, using the Handout for the day. In this phase, relevant parts of movie clips or pop songs were often presented as models. The remaining 70 to 75 minutes were spent on cyclical one-on-one performance testing/coaching using the Personal Card for the day. The 30-plus students formed a circle with their Cards in hand, and the author walked around inside the circle, testing on a one-on-one basis. A student pronounced one item without looking at the Card, and if the targeted sounds were all produced satisfactorily, the author declared a "pass." When one or more features were not acceptable, the author declared a "fail" and quickly pointed out what was wrong (e.g., "Work on the /r/ in 'very'"). One who earned a "pass" put a small circle in the designated box on the Card, while one who failed jotted down the reason for the failure in the box. A student was allowed to try the items only one at a time, whether resulting in a pass or a fail. One who failed in an item tried the same item when subsequent turns came around, until a pass was earned for that item. The number of turns that came around for one student in one session was somewhere between 15 and 30. Near the end of a session, the testing rounds were declared to be over and the Cards were collected. The number of "passes" or circles the student earned was his or her score for the day. The cumulative scores of all the classmates were screen-displayed, from time to time, in the form of a bar chart for everyone to inspect, to nurture friendly competition.

3.3 Recordings

In Session 1 (April 17, 2007) and Session 24 (Feb. 26, 2008), each participant read aloud the following sentence, which was recorded on a PC using Audacity (Mazzoni, 2004).

Thank you very much for everything you did for me.

This particular sentence was chosen because it contained four target consonants (/r/, /f/, /v/, /θ/) and one key vowel (/æ/), apart from the fact that it was a useful formulaic expression worth memorizing The students were told that it was going to be a "pronunciation test" so that they should exhibit their best performance. Before the recording began, the author read it aloud three times as a model, to the whole group. During the recording, the printed sentence was placed in front of them, so they did not need to produce it from memory. Recording was done on a one-on-one basis, with the author operating the PC, in a quiet room. The two large WAV files, one created in Session 1 and the other in Session 24, were separated into 68 files, each containing one utterance by one student without an identification information.

Additional five recordings of the same utterance were made as follows: One recording was of the author's attempt to sound as stereotypically "Japalish" as possible. Segmentally, every English sound was replaced by its closest relative in Japanese phonology (e.g., /θ/ by /s/, /v/ by /b/, etc.), epenthesis was used, and the overall tone was kept as flat as possible. Another was of the author's attempt in the opposite direction – to sound as north-American as possible. This was the same as the pre-recording model utterance students listened to above. The other three recordings were by L1 English speakers of north American origin, two males and one female. All were Japan-based, teaching English to Japanese learners. These recordings were made to encourage judges to use as wide a range on the rating scale (see below) as possible, as well as to elicit baseline data for the human ratings and acoustic analyses below.

All the files containing utterances by the course participants, the author, and the three native speakers were normalized for peak intensity (maximum -3dB), randomized, and allotted serial numbers from 1 to 73.

3.4 Analysis

3.4.1 Perceived Foreign Accent

Perceived degrees of foreign accent were rated by 18 native and near-native speakers. Fifteen of them (age 14-18) were students enrolled in a drama class at an international school in western Japan. They were of diverse nationalities and ethnic origins. Duration of their stay in Japan ranged from three months to 14 years. Asked about their most dominant language, 13 cited English, one Danish, one Italian, and one both English and Japanese. According to their teacher (an L1 English speaker), the Danish student's English is "near fluent," and the Italian student is "strong." Rather than excluding these probably-not-native-enough raters from the beginning based on background information, it was deemed more appropriate to include them in the preliminary Rasch analysis and examine the fit statistics before deciding what finally to do. The other three (age 35-40) were all L1 English speakers from north America. One was the teacher of this drama class, and the other two were teaching English to Japanese learners. Duration of their stay in Japan ranged from 5 to 12 years.

Rating was conducted on a 9-point scale, from 1 "no foreign accent" to 9 "very heavy accent," following Munro and Derwing (1999). The raters were asked to decide on "the overall degree of foreign accentedness" Clarification was made that accentedness concerns both segmental and prosodic features, and emphasis was given that any factors other than pronunciation should be ignored. The raters were informed that they were going to listen to recordings by Japanese learners with several by native speakers interspersed among them. The drama class students and their teacher listened to each file played twice by the author and did the rating collectively. The other two raters did it individually on separate occasions, one in the author's presence and the other by himself.

3.4.2 Perceived Segmental Accuracy

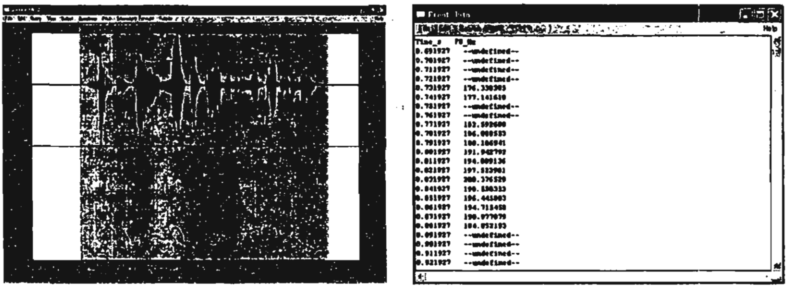
Accuracy of segments was rated by two L1 Japanese English teachers. Rater 1 was the author. The author has an MA in TESOL, a PhD in Applied Linguistics, and a 25-year experience of teaching Japanese learners. Rater 2 was a high school English teacher with a 19-year experience. She is currently conducing an MA research in the field of pronunciation pedagogy.

Rated were the following ten segments: /θ/ and /æ/ in *thank*, /v/ and /r/ in *very*, /f/ in *for*, /v/, /r/, /θ/, and /ŋ/ in *everything*, and /f/ in *for*. These are segments often approximated by Japanese phonemes, hence deemed suitable for examining acquisition or non-acquisition of key English phonemes. These ten segments will be referred to as Items 1 to 10.

The raters played each file multiple times and rated each item on a 3-point scale: 2 ("sufficiently English-like"), 0 ("evidently Japanese"), or 1 ("something in-between" or "not clear hence unable to judge"). After simply confirming these descriptors, there was no further negotiation of the criteria between the raters. They did the ratings independently, each three times, with intervals of three or more days in-between, resulting in six sets of ratings (henceforth referred to as ratings 1A, 1B, 1C, by rater 1; 2A, 2B, 2C, by rater 2). One rating round took approximately one and a half hours.

3.4.3 Pitch

Pitch analysis of each recording was conducted using Praat (Boersma & Weenink, 2005). Praat can return the fundamental frequency (F0) at, by default, every 10 milliseconds in the selected portion of a waveform. The whole waveform of each recording was selected (the left panel, Figure 1), the returned F0 values (the right panel, Figure 1) were copy-pasted onto Excel, and the range and the standard deviation were computed, for each speaker. In so doing, care was taken to watch out for and remove what Ladefoged (2003, p.87-90) calls "micro prosody," or too-sudden pitch changes that are not considered phonologically relevant. Specifically, a pitch change of more than 30 Hz observed across two adjacent 10-millisecond long windows was ignored.



```
Time_s    F0_Hz
0.691927   --undefined--
0.701927   --undefined--
0.711927   --undefined--
0.721927   --undefined--
0.731927  176.330303
0.741927  177.141630
0.751927   --undefined--
0.761927   --undefined--
0.771927  152.592600
0.781927  186.000533
0.791927  188.106941
0.801927  191.942792
0.811927  194.009136
0.821927  197.523901
0.831927  200.376520
0.841927  190.180333
0.851927  196.441003
0.861927  194.715450
0.871927  190.977079
0.881927  186.052193
0.891927   --undefined--
0.901927   --undefined--
0.911927   --undefined--
0.921927   --undefined--
```

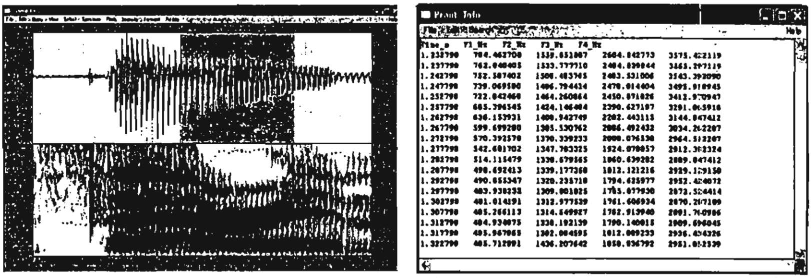
Figure 1. Sample Waveform, Pitch Curve, and Returned F0 Values

3.4.4 /r/ Acceptability

One of the few segmental features for which acoustic quantification is relatively easy is /r/. Ladefoged (2005, p. 55) states that "*whenever* there is an /r/ in a word, the third formant (F3) will be below 2,000 Hz, sometimes falling to as low as 1,500 Hz" (italics mine). This, however, seems an overgeneralization since it is not rare to encounter native speakers whose F3 for an /r/ only goes down to around 2200 Hz (Hagiwara, 1995, as cited in Kent & Read, 2002, p. 181). In fact, one of the three recordings made by native speakers in this study had the lowest F3 of 2297 Hz in *very* and 2259 Hz in *every*.

In the face of such individual variations, this study adopted two approaches to compare the qualities of supposed /r/s in the recordings. (a) If the lowest F3 value of a supposed /r/ was below 2000 Hz, the sound was categorically regarded as acceptable; (b) When one supposed /r/ had a lower F3 than another supposed /r/ *produced by the same speaker*, the former was judged to be *more* acceptable than the latter.

So, the questions were: (a) Were there a greater number of categorically acceptable /r/s in the post-course recordings? (b) Were the supposed /r/s in the post-course recordings more acceptable, i.e., of lower F3 values, than those in the pre-course recordings? To find out, F3 frequencies for supposed /r/s in *very* and *everything* were examined using Praat. On expanded waveforms, the segment corresponding to a supposed /r/ was cursor-selected (the left panel, Figure 2), from which F3 frequencies were obtained (the right panel, Figure 2), and the lowest point was determined.



```
Time_s    F1_Hz       F2_Hz       F3_Hz       F4_Hz
1.232790  784.462700  1535.051807  2604.842773  3575.422119
1.237790  762.040405  1533.777710  2404.839844  3463.297119
1.242790  752.507402  1506.403765  2403.531006  3543.392090
1.247790  739.069500  1496.794434  2470.014404  3495.918945
1.252790  722.042460  1464.260864  2450.071826  3412.970947
1.257790  685.396545  1424.146484  2390.627197  3291.065918
1.262790  636.153931  1400.942749  2202.443115  3144.047412
1.267790  599.699280  1385.530762  2086.492432  3034.242207
1.272790  570.392570  1370.339233  2000.076530  2964.512207
1.277790  542.601702  1347.703325  1924.070057  2912.362324
1.282790  514.115479  1330.679565  1860.639282  2889.047412
1.287790  498.692413  1339.177360  1812.121216  2929.129150
1.292790  490.855347  1320.235710  1794.635977  2952.424072
1.297790  483.930233  1309.001025  1785.077930  2873.524414
1.302790  481.014191  1312.977539  1761.606934  2870.217109
1.307790  485.266113  1314.549927  1782.913940  2891.740986
1.312790  484.934075  1330.192139  1790.140015  2909.696045
1.317790  485.967065  1302.004595  1812.009233  2936.434326
1.322790  485.712891  1436.207642  1850.036792  2951.012339
```

Figure 2. Sample Expanded Waveform, Formants, and Returned F1-F4 Values

4. Results

4.2 Perceived Foreign Accent

Inter-rater correlations among the accent ratings by 18 judges (A-R) are shown in Table 1. Coefficients below 0.6 are underlined. Notably, raters A, C, and M have different rating patterns from the others.

Table 1 Pearson Correlations Between Ratings by the 18 Judges

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
A	-																	
B	0.56	-																
C	0.37	0.26	-															
D	0.41	0.58	0.28	-														
E	0.57	0.67	0.43	0.54	-													
F	0.53	0.76	0.32	0.63	0.61	-												
G	0.59	0.74	0.29	0.59	0.68	0.69	-											
H	0.52	0.70	0.32	0.69	0.61	0.78	0.66	-										
I	0.52	0.67	0.23	0.62	0.57	0.64	0.73	0.72	-									
J	0.41	0.79	0.34	0.64	0.61	0.77	0.64	0.69	0.74	-								
K	0.51	0.71	0.37	0.47	0.66	0.67	0.68	0.55	0.66	0.67	-							
L	0.47	0.76	0.21	0.49	0.57	0.62	0.66	0.60	0.60	0.62	0.72	-						
M	0.59	0.58	0.36	0.32	0.55	0.54	0.53	0.48	0.40	0.52	0.54	0.53	-					
N	0.67	0.75	0.57	0.62	0.70	0.71	0.71	0.75	0.63	0.67	0.63	0.63	0.57	-				
O	0.44	0.71	0.42	0.57	0.67	0.68	0.60	0.65	0.61	0.77	0.63	0.68	0.60	0.69	-			
P	0.55	0.79	0.26	0.61	0.68	0.70	0.79	0.75	0.76	0.73	0.67	0.76	0.52	0.72	0.67	-		
Q	0.67	0.75	0.36	0.60	0.64	0.64	0.76	0.72	0.73	0.64	0.70	0.68	0.48	0.73	0.57	0.77	-	
R	0.40	0.71	0.18	0.56	0.62	0.66	0.70	0.63	0.68	0.73	0.65	0.67	0.52	0.63	0.77	0.69	0.60	-

Winsteps (Linacre, 2005) was run to check, in the Rasch measurement framework (Rasch, 1960; Wright & Stone, 1979), whether these and any other judges had produced ratings with too much "noise" to be meaningfully included in subsequent analyses. (Note that the judges were treated as "items" by Winsteps.) The cutoff was set at 1.30 for both infit and outfit mean squares, following Bond and Fox's (2007; p. 243) guidelines. When misfit judges were identified, they were removed from data and Winsteps was run again. This cycle was iterated until all the raters' infit and outfit mean squares were lower than or equal to 1.30. For this criterion to be met, a total of seven raters (A, C, D, H, I, L, and M) had to be removed through five Winsteps runs, leaving ratings by the other 11 judges to be used for final ability estimation. It turned out that no meaningful relationship was identified between the fit statistics and the background-based native/pseudo-native distinction.

Before presenting further Rasch results, the distribution of mean raw ratings by the finally retained 11 judges is illustrated in Figure 1. The distribution of post-course ratings (M = 5.08, SD = 0.81, Max = 6.45, Min = 3.45) are evidently different from that of pre-course ratings (M = 6.72 , SD = 0.81, Max = 8.18 , Min = 5.09). Be reminded that the higher the number, the heavier the perceived accent was. Incidentally, the author's "Japalish" and "north American" recordings got the mean ratings of 8.64 and 2.00, higher than, and lower than, any recordings by students, respectively. The three recordings by native speakers got mean ratings of 1.18, 1.27, and 1.55, respectively.

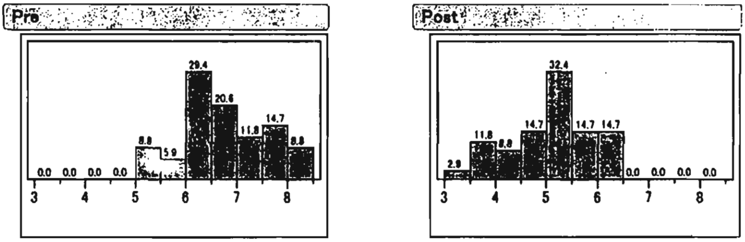


Figure 1 Mean Rating Distributions in Pre- and Post-Course Recordings

Returning to Rasch results, person separation reliability and rater separation reliability were 0.92 and 0.97, respectively, after excluding ratings for four non-student recordings. Hence, person abilities and rater severities turned out to be very reliably diverse. As expected, person abilities were higher for the post-course recordings (M = 0.57, SD =0.71) than for the pre-course recordings (M = -1.10 , SD = 1.01). A paired t test indicated that this was significant, t(33) = -9.51, p =.000, and the effect size was large (d = 1.92).

4.1 Perceived Segmental Accuracy

Inter- and intra-rater correlations among segmental accuracy ratings are illustrated in Table 2. The average correlation among Rater 1's ratings (1A, 1B, and 1C) was 0.868, while that among Rater 2's was somewhat lower at 0.774. The average of all the correlations was 0.753. This was considered an appropriate degree of agreement, neither too high nor too low. Judges giving somewhat different ratings does not necessarily imply less valid ratings than judges agreeing on all ratings (Linacre, 1994, pp. 23-33). In fact, that is exactly the point of having multiple *human* judges rather than a group of clones, in which case only one of them can replace the entire panel of judges. There is even an observation that "when two examiners award different marks, the average is more likely to be correct, or nearly correct, than it is when they award the same mark" (Harper & Misra, 1976, p. 262, as cited in Linacre, 1994, p. 35). Whether ratings in one or more of these rounds were *too* haphazard to be meaningfully included for deriving students' segmental accuracy was an empirical question, to be settled by subsequent Rasch model-data fit analyses.

Table 2 Pearson Correlations Between Ratings in the Six Rounds

	1A	1B	1C	2A	2B	2C
1A	1.000					
1B	0.889	1.000				
1C	0.842	0.869	1.000			
2A	0.624	0.640	0.642	1.000		
2B	0.684	0.706	0.681	0.780	1.000	
2C	0.731	0.753	0.734	0.725	0.810	1.000

Facets (Linacre, 1997) was run to analyze the raw data using the many-facet Rasch analysis framework (Linacre, 1994). The author's first concern was the rater statistics, which is shown in Table 3. It turns out that ratings by Rater 2 were consistently more severe than those by Rater 1, but none of the fit statistics was outside the commonly accepted range of 0.7-1.3 (Bond & Fox, 2007, p. 243). Therefore, the ability results based on all the six rating rounds were adopted.

Table 3 Severity, Standard Error, and Fit Statistics for the Six Ratings

Rating	Measure	*SE*	Infit MS	Infit *Z*	Outfit MS	Outfit *Z*
1A	-0.78	0.08	1.1	1	1.2	1
1B	-0.46	0.07	1.0	0	1.0	0
1C	-0.35	0.07	1.1	1	1.3	2
2A	0.90	0.07	1.0	0	1.0	0
2B	0.56	0.07	0.8	-2	0.8	-1
2C	0.13	0.07	0.9	-1	0.8	-1

Before presenting further Rasch results, raw ratings are summarized in Table 4 for ease of intuitive grasp of the obtained pattern. It can be seen that for every segment, the mean was markedly higher in the post-course recordings.

Table 4 Means and *SD*s of Raw Ratings for Each Segment

		THank	thAnk	Very	veRy	For	eVery	evRy	THing	thinG	For	Total
M	Pre	0.7	1.3	0.4	0.2	0.2	0.4	0.3	0.2	0.3	0.2	4.3
	Post	1.8	2.0	1.8	1.2	1.7	1.6	1.0	1.5	1.5	1.6	15.7
SD	Pre	0.8	0.8	0.7	0.5	0.5	0.6	0.5	0.6	0.6	0.5	4.1
	Post	0.5	0.2	0.4	0.8	0.6	0.6	0.8	0.8	0.7	0.7	3.2

Note. Uppercased parts in the head words correspond to target sounds.

Rasch person separation reliability was 0.99 and item separation reliability was 0.98, which means that both person abilities and item difficulties were very reliably diverse. Just like in the case of holistic ratings above, person measures were higher for the post-course recordings ($M = 1.58$, $SD = 0.74$) than for the pre-course recordings ($M = -1.75$, $SD = 1.36$). A paired t test indicated that this was significant, $t(33) = -14.61$, $p = .000$, and the effect size was large ($d = 3.05$).

Item statistics computed based on all the ratings (i.e., pre-course and post-course recordings combined) are shown in Table 5. Items are tabulated in a difficulty-descending order. Several observations are possible about logit measures and their *SE*s: (a) /r/s were reliably more difficult (i.e., more than 2*SE*s higher) than any other segments; (b) /f/s were reliably more difficult than /v/s; but (c) postvocalic /v/ in *everything* was reliably more difficult than word-initial /v/ in *very*; likewise, (d) postvocalic /θ/ in *everything* was much more difficult than the word-initial /θ/ in *thank*. With regard to fit statistics, /v/ and /r/ in *everything* border on misfit. This probably resulted from the difficulty of reliably judging the qualities of these segments pronounced rather quickly (the portion "*every*" was often pronounced in no more than 250 milliseconds).

Table 5 Item Difficulty, Standard Error, and Fit Statistics for Segmental Analysis

	Measure	*SE*	Infit MS	Infit *Z*	Outfit MS	Outfit *Z*
evRy	1.30	0.09	1.2	3.0	1.9	3.0
veRy	0.96	0.09	1.1	0.0	0.9	0.0
thiNG	0.52	0.09	1.1	1.0	1.2	0.0
THing	0.42	0.09	1.2	2.0	1.0	0.0
For_me	0.35	0.09	0.8	-2.0	0.8	-1.0
For_ev..	0.24	0.09	0.7	-3.0	0.7	-2.0
eVry...	-0.07	0.09	1.0	0.0	1.4	2.0
Very	-0.39	0.10	0.7	-3.0	0.6	-2.0
THank	-0.85	0.10	1.0	0.0	1.0	0.0
thAnk	-2.48	0.11	1.0	0.0	0.7	-1.0

4.3 Pitch

Descriptive statistics for pitch range and pitch *SD* (i.e., *SD* of all the F0s computed in each recording) are given in Table 6. Both pitch ranges and pitch *SD*s were larger in post-course recordings than in pre-course recordings. Paired *t* tests indicated that both differences were significant, *t*(32) = 2.81 , *p* = .008, *t*(32) =3.60 , *p* = .001, respectively. The effect size for pitch range was small (*d* = 0.37), but that for pitch *SD* was medium (*d* = 0.49).

Table 6 · Means and *SD*s of Pitch Range and Pitch *SD* in Pre- and Post-course Recordings

	Pitch Range in Utterance		Pitch *SD* in Utterance	
	M	*SD*	*M*	*SD*
Pre	83.91	36.63	15.11	7.10
Post	102.9	51.90	19.39	7.05

4.4 /r/ Acceptability

The numbers of categorically acceptable /r/s, of which F3s were lower than 2000 Hz, are shown in Table 7. Both in *very* and *every*, there was only one speaker who produced such an /r/ in the pre-course recording, but the number increased to 9 (*very*) and 7 (*every*), respectively, in the post-course recordings. Fisher's exact test indicated that the increase in *very* was significant (*p* = .027) and the effect size was medium (phi = 0.304); the change in *every* failed to reach a significant level (p = .054) and the effect size was small (phi = 0.274).

Table 7 Number of Supposed /r/s with F3 Lower/Higher than 2000 Hz

	veRy		evRy	
	Lower	Higher	Lower	Higher
Pre	1	33	1	33
Post	9	25	7	27

Table 8 Means and *SD*s of F3 (in Hz) of Supposed /r/s in *Very* and *Every*

	veRy		evRy	
	M	*SD*	*M*	*SD*
Pre	2249.0	144.6	2306.9	161.6
Post	2120.9	202.6	2199.5	223.1

Descriptive statistics of F3 frequencies are summarized in Table 8. It can be seen that both in *very* and *every*, the means were lower in the post-course recordings. Paired *t* tests confirmed that the differences were significant both in *very*, *t*(33) = 3.31, *p* =.002, and in *every*, *t*(33) = 3.13 , *p* = .004. These effects were of medium (*d* = -0.61) and small (d = -0.47) sizes, respectively.

4.5 Inter-variable correlations

Finally, correlations between all the variables were checked, the results of which are shown in Table 9. The segmental accuracy measure very strongly correlated (r = .753, p < .01) with the holistic accent measure, accounting for more than 56% of its variance (r^2 = .567). Pitch *SD* also correlated with accentedness significantly (r = .290, p < .05), but the effect size was small (r^2 = .085). Of the two pitch-based variables, *SD* correlated more strongly with accent than the range did. F3 frequencies of /r/ in *very* and *every* also correlated significantly but weakly with accentedness. Of the two /r/-related variables, the continuous F3 frequency correlated more strongly with accent than the 2000 Hz-based dichotomous variable did.

Table 9 Correlations between Examined Variables

	1	2	3	4	5	6	7	8
1. Accent	1.000							
2. Segmental	0.753	1.000						
3. Pitch_Range	0.155	0.195	1.000					
4. Pitch_SD	0.290	0.292	0.854	1.000				
5. veRy_F3	-0.291	-0.369	-0.035	-0.059	1.000			
6. veRy_cat	0.248	0.340	-0.142	-0.052	-0.715	1.000		
7. eveRy_F3	-0.242	-0.273	-0.026	-0.177	0.298	-0.294	1.000	
8. eveRy_cat	0.187	0.221	0.219	0.292	-0.056	0.127	-0.677	1.000

Note. Solid underline: p < .01; Dotted underline: p < .05; "_cat": categorical

Multiple regression was carried out with accentedness as the predicted variables and segmental accuracy, pitch SD, *veRy* F3, and *eveRy* F3 as predictor variables. As can be seen in Table 10, segmental accuracy turned out to be the only significant predictor. Pitch *SD* or F3 frequencies did not have any significant contribution after segmental accuracy was partialed out. R-squared was .57 and adjusted R-squared was .54.

Table 10 Unstandardized and Standardized beta coefficients in Regression Analysis Predicting Accentedness

Variable	*B*	*SE B*	β	*SE* β	*t*-ratio	*p*
Segmental	0.43	0.05	0.72	0.09	7.68	<.0001
Pitch_SD	0.01	0.01	0.07	0.09	0.86	0.39
veRy_F3	-0.00	0.00	-0.01	0.09	-0.14	0.89
eveRy_F3	-0.00	0.00	-0.03	0.08	-0.32	0.75

5. Discussion

The present study compared the same utterance recorded before and after a pronunciation course, and the results were fairly straightforward: the utterance was of markedly better quality after the course. The improvement was identified by human rating as well as by acoustic measurement. The perceived degree of holistic foreign accentedness, the perceived accuracy of key phonetic segments, the range and standard deviation of F0, and the F3 frequency of rhotic /r/s all converged in the same'direction. In plain language, students were able to say "Thank you very much for everything you did for me" using more target-like individual sounds with less flat pitch contour, at the end of a 24-session course than at its beginning.

And how remarkable *is* that? Very. If Japanese junior high school students becoming able to read aloud just one short English sentence with better pronunciation and prosody after one year of instruction does not strike the reader as much of an accomplishment, then he or she should be reminded of what happens, or, rather what does not happen, in Japanese EFL classrooms. The author's 25-year experience of teaching English to Japanese learners at a variety of levels, from junior high school to senior high school to university to graduate programs, as well as his 15-year experience of observing other teachers' English classes across the country, clearly indicates that longer years of learning English in Japan is by no means associated with higher quality of pronunciation. To cite one piece of formal evidence, the author's earlier cross-sectional examination of /r/s produced by high school 1st-, 2nd-, and 3rd-year students revealed no year-group differences (Shizuka, 1993). Few teachers would oppose that, for example, university 1st-year students' average pronunciation of, say, *three* is often no better than junior high school 2nd-year students'. Unlike vocabulary size or syntactic maturity, pronunciation quality does not develop over the years but fossilizes at a very early stage of learning English. Therefore, a large intact group, as opposed to a small selected group, of junior high school students like the one in the present study getting better to a significant and meaningful extent at pronouncing even one sentence containing most of potentially problematic phonetic segments deserves to be regarded as a momentous accomplishment by itself.

In addition, getting better at saying "Thank you very much ..." likely means more than just that. It could also mean getting better at saying other sentences as well. This supposition is endorsed by the author's observation in the course as well as by students' self-reports (Shizuka, in press). They claim that they have become more sound-attentive when orally producing English sentences (the effect size was large, $d = -0.68$) and much better at pronouncing English in general (the effect size was large, $d = -0.78$). Eighty-three percent of the participants answered they have "very much" or "substantially" improved.

If participants of the present study succeeded in upgrading their pronunciation skills, which rarely takes place in ordinary English classes, what was the deciding factor? The author theorizes that it was because the participants were forced to undergo one-on-one pronunciation tests, the results of which accounted for 100% of their course grades. In EFL contexts like Japan, where real needs of English are practically nil outside the classroom, it is a fact of life that students pay attention mainly, if not exclusively, to those aspects of the target language in which they are formally assessed. Vocabulary size and knowledge of grammar, for example, are directly or indirectly assessed in internal tests and external examinations. To state the obvious, when students are given scores or evaluations, they do so on an individual basis. That is, each student receives his or her personal score which specifies his or her position relative to the group (norm-referencing) or some standard (criterion-referencing). In that sense, one can say that, with regard to vocabulary, grammar, etc., students receive teacher-fronted collective instruction *and* individualized evaluation.

On the other hand, the skill, as opposed to the knowledge, of pronunciation seems practically never evaluated. There exist teachers, though not so large in proportion, who do care about their students' pronunciation and exert earnest efforts to make a difference, but reports of positive results are, as long as the author knows, rare at best. Personal communication with such teachers reveal that they depend almost exclusively on teacher-fronted intervention, whereby the teacher tries to correct individual students' pronunciation sporadically or unsystematically while pursuing some other main goal of the class (e.g., reading comprehension of a passage). From an empirical point of view, it is a sheer impossibility to enable everyone in a class of 40 Japanese students to produce an acceptable /r/ just by making them repeat after the model in a collective manner. Accordingly, a predictable end product of such reliance on teacher-fronted instruction is the belief that pronunciation instruction is, after all, a futile endeavor. Such a belief, once adopted, could logically lead to complete negligence of pronunciation teaching.

Which should not happen. As reasoned in the review section above, exactly because it is the age of world Englishes, where nonnative-nonnative interactions proliferate, acquiring good-enough and readily-intelligible pronunciation is becoming more important than ever. The results of the present study combined with those of the self-perception study (Shizuka, in press) strongly imply that individualized evaluation of pronunciation has a potential of achieving what collective teacher-fronted instruction alone has kept failing to achieve. Despite the ingrained belief against "pronunciation policing" seemingly held by some practitioners, pronunciation training, when conducted well, can be a fruitful and rewarding practice, the improvement from which becomes tangible even in the short term. When students sense they have improved, they naturally feel proud of themselves, which motivates them to get even better, feeding into a positive circle. The self-perception study confirmed that participants' open ended verbal comments were of an overwhelmingly positive tone, comprising expressions of their joy, pride, and satisfaction.

The readers who have kept trying to improve their own students' pronunciation with less-than-satisfactory results so far are invited to try out the PCM in their own classrooms. The method is applicable even to situations where the main focus is on something other than pronunciation. In a class of 40 students, having every student produce one short sentence and giving a quick pass/fail feedback with a one-word advice (e.g., "TH in *thank*"), one after another, would be possible within four minutes or so, when carried out efficiently. It follows that eight minutes will give each student two chances to be tested/coached. Saving eight minutes for pronunciation performance tests once, say, every week would sound practical in most classroom situations. To reiterate, the key is to hand out a sheet or card for each student to record his or her performance results and to collect it, which will drive home to the students that how well they do on that test will matter as much as any other written test results will.

Apart from the verification that the course has succeeded in improving participants' pronunciation skills, the present study has made a couple of subsidiary but noteworthy findings. One is that the correlation between nonnative-rated segmental quality and native-judged holistic quality was very strong ($r = .753$). This means more than 56 % ($r^2 = .567$) of the accentedness variance was accounted for by the segmental variance alone. In the multiple regression analysis, nonnative-rated segmental accuracy was the only significant, and very effective, predictor of native-rated holistic accentedness; pitch *SD*, though a significant predictor by itself, did not have a unique contribution beyond segmental accuracy. Although whether segmental or prosodic features have a stronger influence on the intelligibility of the utterances is a question yet to be settled, with evidence and arguments presented favoring either segmental (Jenkins 2000; Riney, Takagi, & Inutsuka, 2005) or suprasegmental features (Tanabe, 2007; Munro & Derwing 1999; Nonaka 2005b; 2005c; Anderson-Hsieh, Johnson, & Koehler, 1992), the data in the present study suggest that segmental accuracy plays a greater (i.e., 50 percent plus) role at least in accentedness perception.

Another observation concerns relative difficulties of consonants for Japanese students. Six rounds of ratings suggest that, among /r/, /f/, /v/, /θ/, the order of difficulty was: /r/ > /f/ > /v/. /θ/ could not be placed easily in the inequality because the word-initial /θ/ in *thank* was easier than /v/ while the word-medial /θ/ in *everything* was much more difficult than /f/. One segment occurring at a comparatively non-salient position such as in the middle of a relatively long word tending to be pronounced less accurately than the same sound occurring at a more salient position such as at the beginning of a word or in the middle of a relatively short word seems to hold true for other sounds as well. The /v/ and the /r/ in *everything* were much more difficult than /v/ and /r/ in *very*, respectively. All these may have been common knowledge among classroom teachers, but it is a contribution of the present study that these patterns have been quantified.

The other findings were about relative usefulness of acoustic variables. First, the *SD* of F0 values in an utterance correlated more strongly with the utterance's perceived accentedness than its range of F0 did. Considering that a range can easily be affected by only one extreme (maximum or minimum) value, it seems reasonable that an *SD* more appropriately reflects the dispersion of fundamental frequencies in one utterance. This result implies that as an index of voice pitch variability, the *SD* may be preferred over the range or the maximum value, the two indexes adopted by Yabuuchi'and Satoi (2001). Second, the F3 frequencies of /r/ correlated more strongly, than the 1/0 values based on 2000 Hz cutoff did, not only with overall segmental accuracy but also with perceived accentedness. The implication is that F3-based quality judgment of a supposed /r/ had better be made according to the degree of F3 lowering than as occurrence/non-occurrence categorization in relation to a fixed value of 2000 Hz. This is consistent with Hagiwara's suggestion (cited in Kent & Read, 2002, p. 181) that the extent of F3 lowering is best determined in relation to an intra-personal neutral value of F3 rather than in relation to some speaker-independent critical frequency value.

Conclusion

The present study has verified that a pronunciation course did produce expected results of improving phonological qualities of participants' scripted utterance, and has presented an argument that the effect should be primarily interpreted as positive backwash of continuous, one-on-one, performance testing. Before concluding, limitations of this study need be specified and future direction suggested. Having said that getting better at pronouncing one formulaic utterance likely means getting also better at pronouncing other utterances, it is obviously necessary to carry out future research to substantiate this supposition. One direction is to use less controlled tasks where the amount of attention payable to pronunciation is smaller. Having students produce some unscripted speech would be a promising possibility. With regard to acoustic analysis, the present study was able to examine only one aspect of prosody: vocal pitch. A common definition of prosody includes intensity and duration in addition to pitch (Kent & Read, 2002). Although intensity is often not regarded as a useful variable to measure (Ladefoged, 2003, p. 93), duration surely seems worth exploring. In fact, the author noted that the first syllable of *very* in post-course recordings tended to be noticeably longer than that in pre-course recordings. Probably, that was one factor contributing to perceived better quality of the post-course utterances. The duration of that particular syllable was not measured in this study due to the difficulty of reliably determining its syllable boundaries even on an expanded waveform, but exploration in that direction using different speech samples might prove productive.

Acknowledgment

The author thanks Ms. Hiroko Sato, a budding researcher and seasoned EFL practitioner, for cooperating in the arduous segmental rating task. He also expresses gratitude for anonymous native speaker raters who participated in accentedness rating. Finally, special thanks go to Professor Nobuo Yuzawa at Takasaki City University of Economics

for his valuable advice on interpreting F3 values in /ɪ/.

References

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42,* 529-555.

Arimoto, J. (2005). Hatuon shido ni okeru kyoshi no yakuwari [The teacher's role in pronunciation teaching]. *The English Teachers' Magazine, 54,* 27-29.

Asami, M. & Tanaka, A. (2005). Juu dankai hatuon shido wo kensho suru [Verifying the effects of the 10-step pronunciation teaching]. *The English Teachers' Magazine, 54,* 31-33.

Boersma, P. & Weenink, D. (2005). Praat: doing phonetics by computer (Version 4.6.12) [Computer program]. Retrieved July 27, 2007, from http://www.praat.org/

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model.* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Crystal, D. (2003). *English as a global language.* (2nd ed.). Cambridge: Cambridge University Press.

Esling, J. H. & Wong, R. F. (1983). Voice quality setting and the teaching of pronunciation. *TESOL Quarterly, 17,* 89-95.

Flege, J. E. (1980). Phonetic approximation in second language acquisition. *Language Learning, 30,* 117-134.

Flege, J. E. (1981). The phonological basis of foreign accent: A hypothesis. *TESOL Quarterly, 15,* 443-455.

Flege, J. E., Takagi, N., & Man, V. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Language and Speech, 38,* 25-55.

Hagiwara, R.(1995). *Acoustic realization of American /R/ as produced by women and men.* (University of California-Los Angeles Working Papers in Phonetics No. 90). Phonetics Laboratory, University of California-Los Angeles.

Harper, A. E. Jr., & Misra, V. S. (1976). *Research in examinations in India.* New Delhi, India: National Council of Educational Research and Training.

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

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

Kent, R. D. and Read C. (2002). *The acoustic analysis of speech.* (2nd ed.). Albany, NY: Singular/Thomson Learning.

Kosuge, K. (2005). Kihon teki na shidoho ni atarashii kuhu wo kuwaete [Add'new ideas to traditional teaching techniques]. *The English Teachers' Magazine, 54,* 11-13.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques.* Malden, MA: Blackwell.

Ladefoged, P. (2005). *Vowels and consonants: An introduction to sounds of languages.* (2nd ed.). Malden, MA: Blackwell.

Linacre, J. M. (1994). *Many-facet Rasch measurement..* (2nd ed.). Chicago, IL: MESA Press.

Linacre, J. M. (1997). Facets (Version 3.0) [Computer software]. http://www.winsteps.com/

Linacre, J. M. (2005). Winsteps (Version 3.55) [Computer software]. http://www.winsteps.com/

Makino, T. (2005). Naze seito wa hatuon wo machigau ka [Why do students produce wrong sounds?]. *The English Teachers' Magazine, 54,* 15-17.

Mazzoni, D. (2004). Audacity (Version 1.2.1) [Computer software] Retrieved June 15, 2006, from http://audacity.sourceforge.net/

Munro, M. J. & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49,* 285-310.

Nakanishi, N. (2004). Japalish hatuon ga imi rikai to taijin miryoku ni oyobosu eikyou [The effects of Japalish pronunciation on comprehension and interpersonal attraction]. Proceedings of the 7th KELES Graduation and Masters Theses Presentation Seminar, 51-56.

Nonaka, I. (2005a). *Eigo jita no tukuri kata [How to pronounce English right].* Tokyo: Kenkyusha.

Nonaka, I. (2005b). Eigo rashiku kikoeru hatuon no kotu wa? [What is the knack for producing English-like sounds?]. *The English Teachers' Magazine, 54,* 14.

Nonaka, I. (2005c). Motto purosodii wo [Teach more about prosody!]. *The English Teachers' Magazine, 54,* 19-21.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmark Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago).

Riney, T. J., Takagi, N. & Inutsuka, K. (2005). Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly, 39,* 441-466.

Shizuka, T. (1993). Task variation and accuracy predictor in interlanguage phonology production.. *Kanto-koshin-etsu Association of Teachers of English Bulletin, 7,* 63-80.

Shizuka, T. (1995). Kojin kaaddo hoshiki tango risuto hatuon shidou no kouka ni kansuru jissho teki kenkyuu [Effects of Personal Card Approach to pronunciaticn teaching]. *Kanto-koshin-etsu Association of Teachers of English Bulletin, 9*, 11-19.

Shizuka, T. (in press). The effects of a 24-session EFL pronunciation course reflected in learners' self-reports. *JACET Journal.*

Smith, L. E. (1983). *Readings in English as an international language*. Pergamon Press.

Suter, R. W. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning, 26*, 233-253.

Tanabe, Y. (2007). Questioning commonly held beliefs in pronunciation teaching [Hatuon shido minaoshi ron kara: Onsei komyunikeshon shido no joushiki wo kensho suru]. *The English Teachers' Magazine, 56*, 14-17.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Yabuuchi, S. & Satoi, H. (2001). Prosodic characteristics of Japanese EFL learners' oral reading comparison between good and poor readers. *Language Education and Technology, 38*, 99-112.

Yamada, T., Adachi, T., & ATR Institute (1999). *Eigo supiikingu kagaku teki jotatu ho [A scientific method of improving your English speaking skills]*. Tokyo: Kodansha.