Effects of Listening Free Written Recall Tasks on Development of L2 Listening Comprehension Ability

Hideki SAKAI: Language Education

Key words: listening instruction, free written recall tasks, listening comprehension

1. Introduction

Recently, teaching listening in English has drawn much attention from junior high and senior high school teachers in Japan. One reason may be that the focus of instruction as stipulated in the official Course of Study has been shifted toward development of oral communicative skills. Another probable reason is that the national center examination for universities just started to administer listening tests as of January, 2006. As a washback effect of the inclusion of listening tests in the national center examination, senior high school teachers have felt increased pressure to teach listening. However, in most classrooms, teachers play recordings of English, tell students to answer comprehension questions, and check their answers. In other words, listening is still "not being taught but tested" (Sheerin, 1987, p. 126). Thus, how listening should be taught has become one of the most interesting issues in English education in Japan.

Although several methods for listening instruction have been proposed (e.g., Flowerdew & Miller, 2005; Harmer, 2001; Peterson, 2001; Ur, 1984), only a handful of researchers have attempted to study the effects of listening instruction on the development of listening comprehension ability (El-Koumy, 2000; Mizohata, 2003; Onaha, 2004; Sugita, 2005; Tamai, 1992). Most of these studies found positive effects from various listening instruction methods. However, their results must be interpreted with caution because of the lack of a control group in the research design, misuse of statistical analysis, and the limited number of testing instruments.

Listening comprehension is considered to be complex in that several processes take place simultaneously (Buck, 2001; Morley, 2001; Peterson, 2001). For example, Buck (2001, pp. 15-21) listed the following processes: (a) understanding words (recognizing words and understanding their meanings), (b) processing idea units (taking the meanings of individual words, and combining them to construct the meaning of complete utterances), (c) processing connected discourse, and (d) using world knowledge (making inferences). Following this view of listening, listening instruction needs to provide learners with opportunities to use a variety of processes. In addition, listening comprehension ability should be assessed from different angles.

Shadowing and dictation tasks, on which most of the previous studies focused, may require learners to carry out a limited number of processes. Buck (2001) pointed out that learners may just need the skill of word recognition for dictation tasks (p. 77) and that in shadowing tasks, learners are able to recognize and repeat sounds without processing the meaning (p. 79). In terms of Buck's list of listening processes, shadowing and dictation tasks may provide opportunities only to practice the process of understanding words. For this study, I focused on listening free written recall tasks (hereafter, listening recall tasks) in which, after listening to the passage, learners are asked to write

down everything they remember from it. Several studies (Sakai, 2005; Sakai & Tarui, 2006) have demonstrated that listening recall tasks require learners to carry out a variety of listening processes such as retrieving vocabulary items based on perceived sound clues, understanding idea units, understanding the connection of the discourse, and making use of inference. Thus, as James (1986) argued, listening recall tasks can be used as a teaching technique.

In this study, I attempted to investigate the effects of listening recall tasks on the improvement of listening comprehension ability with a pretest-posttest, quasi-experimental design with a control group, using three types of listening tests which were intended to measure different aspects of listening comprehension ability. This study also aimed at checking the necessity of using three types of listening tests as measurement instruments.

2. Empirical Studies of Effects of Listening Instruction

So far, only a few studies have investigated the effects of listening instruction over a period of time on the development of listening ability. The listening instructions targeted were shadowing (Tamai, 1992), a combination of shadowing and dictation (Onaha, 2004), listening strategy training (Mizohata, 2003), skills-based instruction and whole-language instruction (El-Koumy, 2000), and phrasal listening instruction (Sugita, 2005). I will review these studies in this section and point out some problems with them.

Tamai (1992) compared shadowing, which he called *follow-up*, with dictation to examine the effects of shadowing on listening ability development. Ninety-four high school students were divided into two groups: an experimental group which received shadowing instruction ($n = 47$) and a control group which received dictation instruction ($n = 47$). Both groups received a total of 13 listening instruction lessons once a week for three and a half months. Secondary level English proficiency tests (SLEP) were used as a pretest and a posttest. According to Tamai (1992, pp. 57-58), the results showed that the gain of the experimental group was significant, $t(92) = 6.82$, $p = .01$, whereas the gain of the control group was not significant, $t(92) = 2.11$, $p = .05$ [sic]. In addition, he reported that although the difference of the pretest between the groups was not significant, $t(92) = .97$, $p = .005$ [sic], the difference of the posttest between the two groups was significant, $t(92) = 2.11$, $p = .05$. He suggested that shadowing may be a more effective listening instruction technique than dictation (p. 59). It seems that his presentation of statistical values and some of his interpretations of the statistics were wrong or ambiguous; for example, he reported no significant differences, showing a p-value of .005. Above all, he did not control the significance level to avoid Type I error. Because he used four t-tests, the alpha level should have been set at a .012 level by the Bonferroni adjustment. With this adjusted alpha level, only the gain of the experimental group was significant; however, there were no statistically significant differences between the experimental group and the control group. In other words, his results suggested that shadowing may be effective in improving listening ability, but it was not clear that shadowing was more effective than dictation.

Onaha (2004) reported on two studies which examined the effects of the combined practice of shadowing and dictation on improvement of listening comprehension ability. Japanese university students ($n = 43$ for her first study and $n = 62$ for her second study) continued shadowing and dictation

in class and at home during a period of one semester. They took listening comprehension tests (NHK listening tests) as a pretest and a posttest. Paired-samples *t*-tests showed statistically significant differences between the pretest and the posttest in both studies, suggesting that her participants improved their listening ability. She also examined changes of memory ability and discussed the relationship between the development of listening ability and memory ability. Although she found statistically significant improvement in listening ability, one must be cautious in attributing the improvement only to the treatment because her research design did not include a control group. The improvement may have been facilitated by other factors, for example, maturation (Brown, 1988, p. 32).

Mizohata (2003) provided 16 sessions of listening strategy training to 167 Japanese high school students for about 3 months. Two versions of the third grade English proficiency tests of the Society for Testing English Proficiency (STEP) were administered as a pretest and a posttest respectively. That is, his pretest was different from his posttest. Based on their scores on the pretest, the students were divided into three groups: low level, mid level, and high level. The results showed that although students at the mid and high levels did not improve their listening ability, students at the low level showed statistically significant improvement after receiving training in listening strategy. He discussed the results in terms of individual differences such as learning strategies and learning style preferences.

El-Koumy (2000) compared the effects of skills-based instruction and whole-language instruction on listening comprehension. Ninety-six Egyptian university students learning English as a foreign language (EFL) were assigned to one of two groups: the skills-based group and the whole-language group. Each group was further divided into two groups based on their scores on the pretest. The skills-based group received instruction in 15 micro-skills such as "identifying isolated speech sounds" and "identifying stressed syllables in individual words." The whole-language group was provided instruction with its focus placed on the four language skills (speaking, listening, reading, and writing). Both groups received instruction once a week for 5 months and took a listening comprehension test from the TOEFL at the end of the study as a posttest. The results showed that, for the low level, there were no statistically significant differences between the skills-based group and the whole-language group, whereas for the high level, significant differences were found between the two groups, $t = 2.92$, $p < .01$. Therefore, El-Koumy suggested that skills-based instruction does not lead to improvement of listening comprehension and that whole-language instruction may not be effective if students do not have basic skills. Thus, a combination of skills-based instruction and whole-language instruction was argued to be more effective.

Sugita (2005) investigated the effects of phrasal listening instruction with 48 Japanese college students. In five phrasal listening instruction sessions, students listened to a passage with pauses inserted and were told to think during the pauses about the meaning of the English sentences they had just heard. Sugita made a listening test based on the pre-2nd STEP test and G-TELP as the pretest and posttest. The students were divided into three groups based on their scores on the pretest: advanced group ($n = 12$), intermediate group ($n = 23$), and lower group ($n = 13$). Results showed that a statistically significant tendency was found in the interaction effect of the test factor and the listening proficiency factor ($p < .10$). He further performed an analysis of simple effects according to each level

and found that the differences between the pretest and the posttest were statistically significant at the 1% level for the intermediate group and the lower group and that the probability for the advanced group was less than 10%. Thus, Sugita concluded that phrasal listening instruction was effective, and at the same time, the students' proficiency level influenced the effectiveness of the instruction.

Several points should be made here. First, these researchers' results are not conclusive because of their research design flaws. Three studies (Mizohata, 2003; Onaha, 2004; Sugita) reported the general effectiveness of targeted listening instruction. However, because they did not include control groups, it is not still clear whether the instruction is a causal factor for the improvement of listening ability. One study (Tamai, 1992) argued for the comparative effects of shadowing to dictation. However, reexamination of the results indicated that both techniques may have been effective for the development of listening ability. Second, all the studies used listening tests developed by outside institutions, but did not report the reliability or validity of the test instruments except for Sugita (2005). Sugita (2005) argued that the listening tests of STEP and G-TELP were appropriate as a measurement of listening comprehension test for his study and reported the reliability of the tests. However, the reliability coefficients of the pretest and the posttest were quite low (.46 and .61, respectively). He admitted this problem, but went on to further analyses. No other studies mentioned the validity or reliability of the test instruments. Furthermore, each study utilized only one type of listening test. If listening comprehension does indeed involve complex processes, it should be tested using different instruments. Lastly, as far as I know, no researcher has studied the effects of listening recall tasks, on which this study focused.

Taking the problems found in the previous studies into consideration, I set up a control group and used three types of listening tests which were intended to measure different aspects of listening comprehension ability. I formulated two research questions: Will different types of listening tests measure different aspects of listening comprehension ability; and are the listening recall tasks effective for improvement of second language (L2) listening ability, and if so, what aspects of listening ability will improve through the listening recall tasks?

3. Method

3.1 Participants

The study began with 28 participants, all of whom were Japanese speaking university students enrolled in either the Teaching English Oral Communication Skills II class or the Theories for Teaching EFL II class in the fall semester of 2005 at Shinshu University. Both courses provided elective credits for teachers' certificates. Thus, students in both courses were similar in terms of their interest in English, their English ability, and their class schedules (i.e., the amount of the English study per week). Some participants took both courses, whereas others took one of the two courses. Participants ($n = 17$) enrolled in the Teaching English Oral Communication Skills II class were assigned to the listening recall tasks group (LG), whereas participants ($n = 11$) who only took the Teaching EFL II class were assigned as the control group (CG). Because some participants did not complete all the tests and treatments or did not agree to sign the consent form, the final number of the participants in each group was ten. The LG consisted of 10 female students (9 third-year students and

1 fourth year student), while the CG consisted of 1 male fourth year student and 9 female students (8 third-year students and 1 fourth year student).

Of the previous studies mentioned above, three studies (El-Koumy, 2000; Mizahata, 2003; Sugita, 2005) found listening proficiency to be one influential factor for the effects of listening instruction. However, this factor was not considered for this study mainly because the number of the participants for this study was too small for the further division of the participants based on their listening proficiency level.

3.2 Experimental Design

This study was a pretest-posttest design with a control group. The LG received six sessions of listening practice for 2 months. Before and after the treatment, the LG took three types of listening tests as a pretest and a posttest. The CG took the listening tests at about 3-month intervals. Table 1 shows the experimental schedule.

Table 1

Experimental Schedule

	LG	CG
Pretest (Tests 1, 2, & 3)	October 4 (Test 1) and 11 (Tests 2 & 3)	October 25 or 26
Treatment 1	October 11	
Treatment 2	October 18	
Treatment 3	October 25	
Treatment 4	November 8	
Treatment 5	November 22	
Treatment 6	November 29	
Posttest (Tests 1, 2, & 3)	December 13	January 17, 18, or 19

3.3 Treatment

For the LG, each listening practice session lasted about 15 minutes and consisted of six phases: (a) a review of the previous practice, (b) the first listening, (c) the second listening, (d) self-scoring, (e) listening with the script, and (f) reflection.

In the listening phases, participants listened to a passage read in English and were told to write down in English everything that they understood. While listening, they were not allowed to take notes, but were asked to try to simply comprehend the passage. Immediately after the listening passage was over, I told the participants to start writing. For the second listening, I emphasized that it was important for them to listen again as if they were listening to the passage for the first time. This instruction was provided because in the second listening practice of the first session on October 11, 2005, I found some participants trying to listen only to parts which they had not understood. In the phase of self-scoring, participants were told to read the Japanese translations and to score their second recall, (b) to read the script, (c) to identify the parts which they had misunderstood, and (d) to consult the dictionary if they had unknown words. Then, participants had another chance to listen to the

passage while looking at the script. Lastly, participants were told to reflect on their misunderstanding, difficulties, and possible reasons for misunderstanding. I collected the worksheet. In the next listening practice session, I gave the sheets back and explained some points about their listening difficulties. Then, participants listened to the passage again. Thus, participants had opportunities to listen to each passage four times in total.

The listening passages used in the listening practice were selected from Obunsha (2004a, 2004b, 2004c) which contained the STEP tests (see Appendix A). Most of the passages were at the pre-2nd grade level, which I considered to be the appropriate level for the participants.

3.4 Testing Instruments

Three types of listening tests were selected because they were supposed to measure several aspects of listening ability. Test 1 was the Oxford Placement Test 2 (Allen, 1992) which originally contained 100 items. Participants were asked to listen to a sentence (e.g., *What do you think of the Bell School T-shirts? I really like them.*), to read the printed sentence (e.g, *What do you think of the Bell School [teachers/T-shirts]? I really like them.*), and to select the word which they heard. This test can be considered to be a type of phonetic discrimination task (Buck, 2001, p. 63). Thus, this test aims at measuring the ability to distinguish similar phonemes. Item analyses were conducted on the results of the pretest. Fifty-nine of the corrected item-total correlations were found to be negative. Thus, these items were excluded from the following analyses. The coefficient alpha for Test 1 with 41 items was .77.

Test 2 was a discourse completion test from TOEIC Part 2 (Saegusa, 1998, pp. 23-24) which contained 30 items. Participants were asked to select the best response out of the three choices following a question. For example, participants listened to a question *Where are we going to park?* and three choices: (a) *This park is filled with old trees.*, (b) *Then, let's go to the park on foot.*, and (c) *The lot on Second usually isn't full.* Then, they were asked to select the most appropriate response. This test can be categorized as response evaluation (Buck, 2001, p. 64). This test requires test-takers to process idea units, that is, the ability of "taking the meaning of individual words, and combining these together to construct the meaning of complete utterances" (Buck, 2001, p. 16). In the above example, if test-takers understand the literal meaning of the interrogative word *where* and the verb *park*, they can come to the correct answer. Because the questions in Test 2 consisted of short single utterances, the processing of connected discourse to grasp "a semantic relation between one element in a text and another element" (Buck, 2001, p. 17) seems to be less involved in this test. This ability to process connected discourse was dealt with in Test 3. Item analyses on the results of the pretest showed that 14 items yielded negative corrected item-total correlations. These 14 items were excluded from the analyses. The coefficient alpha for Test 2 with 16 items was .70.

Test 3 was derived from the STEP 2nd grade listening comprehension test (Obunsha, 2002, pp. 48-51) with 20 items. Participants were told to listen to passages and a question and to choose the best answer of the four choices which were printed on the sheet. The following is an example from the test:

[passage] Nancy had long hair for many years. She liked the way it looked, but it was too hot during the summer. This summer, while on vacation at the beach, she finally decided to get her hair cut. She really likes her new hairstyle, but her

friends and family think that she looked better with long hair.

[question] Why did Nancy decide to get a haircut?

[printed choices] 1. She didn't like her old hairstyle.

2. She wanted to look like her friends.

3. Her long hair was hot in the summer.

4. Her family asked her to change her hairstyle.

In order to come to the correct answer, test-takers should understand the meaning of individual words and phrases of the first three sentences and grasp the logic that because it was too hot during the summer, she decided to have her hair cut. Thus, this test aims at assessing the ability to process idea units and connected discourse. Item analyses on the scores of the pretest revealed that 2 of the 20 corrected item-total correlations yielded negative correlations. Therefore, the two items were excluded from the analyses. The coefficient alpha for Test 3 with 18 items was .75.

3.5 Statistical Analyses

To answer the first research question, three correlation coefficients were calculated among the three tests. For this analysis, the scores of pretests were used with the two groups combined as one group. Using the Bonferroni approach to control Type I error a p value of less than .017, obtained by dividing .05 by 3, was required for significance.

To answer the second research question, a one-way multivariate analysis of variance (MANOVA) was conducted on the gain scores computed by subtracting the pretest score from the posttest score.

3.6 Hypotheses

Based on the two research questions, I set four hypotheses for this study as follows:

Hypothesis 1: There will be no statistically significant correlations among the three types of listening tests.

Hypothesis 2: The listening recall group will improve their phonetic discrimination skills more than the control group.

Hypothesis 3: The listening recall group will improve their ability to understand the details of a passage more than the control group.

Hypothesis 4: The listening recall group will improve their ability to understand the discourse connection more than the control group.

4. Results

4.1 Correlational Analysis

Before the correlational analysis, I checked the normality of the distributions of scores in the three tests and the linear relationship among the scores of the three tests. Table 2 shows the descriptive statistics of the three tests administered before the treatment. The values of $z$skewness and $z$kurtosis for the three tests fell with in the range of ± 1.96. Thus, the three distributions were not statistically significant different from the normal distribution in terms of skewness and kurtosis. I also checked the $z$scores of the three tests and found no outliers in the scores. The inspection of the

scatter graphs of the scores did not indicate non-linear relationships among the three tests. Then, I computed correlation coefficients (Table 3).

Table 2

Descriptive Statistics for the Three Listening Tests (N = 20)

	Test 1	Test 2	Test 3
M	28.25	7.65	13.35
95% CI			
Lower limit	25.87	6.15	11.76
Upper limit	30.63	9.15	14.94
SD	5.08	3.20	3.41
Skewness	-0.45	0.51	-0.92
SE of skewness	0.51	0.51	0.51
Kurtosis	-0.31	0.11	0.50
SE of kurtosis	0.99	0.99	0.99

Note. CI = confidence interval.

Table 3

Intercorrelations for the Three Pretests (N = 20)

Tests	1	2	3
1. Test 1	---		
2. Test 2	.13	---	
3. Test 3	.14	.27	---

Note. None of the coefficients was significant.

There were no statistically significant correlations among the three tests. Although the correlation between Test 2 and Test 3 was slightly higher (r = .27) than that of Test 1 and Test 2 (r = .13) or that of Test 1 and Test 3 (r = .14), all the correlation coefficients were small --- less than .30 (Green & Salkind, 2005, p. 256). Thus, the first hypothesis about the correlations among the three test scores was confirmed, suggesting that the three tests may measure different aspects of listening comprehension ability.

4.2 Gain Analysis

Table 4 indicates the descriptive statistics for the two groups. For Test 1, the LG improved from the pretest to the posttest by 2.20 on average, whereas the mean gain score for the CG was 0.70. For Test 2, the CG improved more than the LG (1.60 vs. 0.60). For Test 3, the mean gain scores of the LG and the CG were almost the same (0.80 vs 1.00).

Table 4

Descriptive Statistics for the Three Listening Tests by Tests and Time

	LG			CG		
Listening tests	Pretest	Posttest	Gain	Pretest	Posttest	Gain
Test 1						
M	28.10	30.30	2.20	28.40	29.10	0.70
95% CI						
Lower limit	24.32	26.94	0.66	24.72	26.68	-2.04
Upper limit	31.88	33.66	3.74	32.08	31.52	3.44
SD	5.28	4.69	2.15	5.15	3.38	3.83
Skewness	-0.75	-0.29	-0.57	-0.22	0.35	0.26
SE of skewness	0.69	0.69	0.69	0.69	0.69	0.69
Kurtosis	-0.23	-1.05	0.37	0.22	-1.05	0.08
SE of kurtosis	1.33	1.33	1.33	1.33	1.33	1.33
Test 2						
M	8.00	8.60	0.60	7.30	8.90	1.60
95% CI						
Lower limit	5.92	6.75	-1.97	4.73	7.38	-0.56
Upper limit	10.08	10.45	3.17	9.87	10.42	3.76
SD	2.91	2.59	3.60	3.59	2.13	3.03
Skewness	0.48	-1.08	-0.38	0.74	-0.27	-1.39
SE of skewness	0.69	0.69	0.69	0.69	0.69	0.69
Kurtosis	-1.83	1.50	-0.21	-1.83	-0.33	1.40
SE of kurtosis	1.33	1.33	1.33	1.33	1.33	1.33
Test 3						
M	13.20	14.00	0.80	13.50	14.50	1.00
95% CI						
Lower limit	10.73	12.00	-0.50	10.97	12.91	-0.26
Upper limit	15.67	16.00	2.10	16.03	16.09	2.26
SD	3.46	2.79	1.81	3.54	2.22	1.76
Skewness	-0.78	-0.15	-0.19	-1.24	-1.02	0.00
SE of skewness	0.69	0.69	0.69	0.69	0.69	0.69
Kurtosis	1.09	-0.45	-1.59	1.31	0.25	0.08
SE of kurtosis	1.33	1.33	1.33	1.33	1.33	1.33

Note. CI = confidence interval.

Before performing a one-way MANOVA, I checked the normality of the distributions of the gain scores of the two groups in the three tests. The values of $z$skewness and $z$kurtosis for the three tests fell with in the range of ± 1.96 except for the $z$skewness for Test 2 of the CG (skewness = -1.39, SE of skewness = 0.69, $z$skewness = -2.01). The distribution of gain scores of Test 2 of the CG was negatively

skewed; in other words, it had more scores above the mean. Thus, the other five distributions were not statistically significantly different from the normal distribution in terms of skewness and kurtosis. I also checked the z scores of the gain scores of the three tests of the two groups and found no outliers in the scores.

A one-way MANOVA found no statistically significant differences among the groups on the dependent variables, $F(3, 16) = .80$, $p = .51$, Wilks's $\Lambda = .87$. The multivariate η^2 was .13. Therefore, the three hypotheses were not confirmed, suggesting that the listening recall tasks were not so effective in improving listening ability as measured by the three tests as they were hypothesized to be.

5. Discussion

5.1 Research Question 1: Will Different Types of Listening Tests Measure Different Aspects of Listening Comprehension Ability?

According to the correlational analysis, the three tests yielded statistically non-significant and low coefficients (.13 to .27). The results suggested that the three listening tests may measure different aspects of listening comprehension ability and lent support to the supposition mentioned in the section on test instruments that Tests 1 to 3 may measure the ability to distinguish similar phonemes, the ability to process the idea units, and the ability to process connected discourse respectively. The results also suggested that listening comprehension may be better understood not as a unitary construct but as a construct involving multiple processes. If so, arguably, studies which examine the effects of listening instruction on the development of listening ability need to specify what aspects of listening comprehension ability will be tested and to use appropriate listening tests to measure those aspects of listening comprehension ability.

5.2 Research Question 2: Are The Listening Recall Tasks Effective for Improvement of L2 Listening Ability, and If So, What Aspects of Listening Ability Will Improve Through the Listening Recall Tasks?

At the first sight of the descriptive statistics, the LG seemed to have improved in the test of phonetic discrimination (Test 1) more than the CG; on the other hand, the CG improved in the test of idea unit processing (Test 2) more than the LG. Statistical analyses showed no statistically significant differences between the two groups in the three tests. The results suggested that listening recall tasks may not be effective in improving listening ability.

Because the number of participants was small, there might be Type II error. I would like to discuss two other possible reasons why there were no statistically significant differences in the three tests between the LG and the CG. First, the results may have been due to the short duration and small number of instruction sessions. In this study, listening practice was provided six times over a period of about 2 months. Sugita (2005) designed the treatment which was the shortest and the most infrequent among the previous studies. Although the number of listening practice sessions (5 times) was less than in this study (6 times), his participants had a chance to listen to one passage more times (8 times) than in this study (4 times); thus, his participants had more exposure to English than the participants in this study. The other studies provided a more prolonged period of instruction. Therefore, it may be possible that if participants have listening practice through listening recall tasks for a longer time and

more frequently, their listening ability will improve.

The second possible reason for there being no statistically significant differences between the LG and the CG may be the effect of learning opportunities outside the study. The improvement of the CG in Test 2 was greater than that of the LG. Participants for this study were university students who wanted to get a teaching certificate for English. They were taking other courses related to English education, literature, and grammar during the semester. Thus, they were assumed to have learning opportunities outside the study as well. These chances to use English elsewhere may have influenced the results of this study.

This quasi-experimental study showed the necessity for use of different types of listening tests, but did not clarify the effects of listening recall tasks on the development of listening comprehension ability. In the future, it will be necessary to make listening instruction longer and more frequent and to control participants' exposure to English outside the study.

References

Brown, J. D. (1988). *Understanding research in second language learning.* Cambridge: Cambridge University Press.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

El-Koumy, A. S. A. (2000). *Effects of skills-based versus whole language approach on the comprehension of EFL students with low and high listening ability levels* (ERIC Document Reproduction Service No. ED449670).

Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice.* New York: Cambridge University Press.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS for Windows and Machintosh: Analyzing and understanding data* (4th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Harmer, J. (2001). *The practice of English language teaching* (3rd ed.). Harlow, Essex: Pearson Education Limited.

James, C. J. (1986). Listening and learning: Protocols and processes. In W. H. Bartz, & J. B. Goepper (Eds.), *Second language acquisition: Preparing for tomorrow* (pp. 38-48).

Mizohata, Y. (2003). Listening strategy training for EFL learners with different learning styles. *Language Education & Technology, 40,* 35-60.

Morley, J. (2001). Aural comprehension instruction: Principles and practices. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed.) (pp. 69-85). Boston, MA: Heinle & Heinle.

Obunsha. (2002). *Eiken 2 kyu zenmondaishu.* Tokyo: Obunsha.

Obunsha. (2004a). *Eiken 4 kyu zenmondaishu.* Tokyo: Obunsha.

Obunsha. (2004b). *Eiken 3 kyu zenmondaishu.* Tokyo: Obunsha.

Obunsha. (2004c). *Eiken jun 2 kyu zenmondaishu.* Tokyo: Obunsha.

Onaha, H. (2004). Effect of shadowing and dictation on listening comprehension ability of Japanese EFL learners based on the theory of working memory. *JACET Bulletin, 39,* 137-148.

Peterson, P. W. (2001). Skills and strategies for proficient listening. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed.) (pp. 87-100). Boston, MA: Heinle & Heinle.

Saegusa, Y. (1998). *TOEIC® test kanzen mogi mondaishu.* Tokyo: The Japan Times.

Sakai, H. (2005). An examination of free written recall tasks as listening comprehension tests. *JABAET (The Japan-Britain Association for English Teaching) Journal, 9,* 49-62.

Sakai, H., & Tarui, C. (2006). Free written recall tasks as L2 listening comprehension tests: Based on Japanese high school

students' performance. *The Bulletin of The Chubu English Language Education Society, 35,* 39-46.

Sheering, S. (1987). Listening comprehension: Teaching or testing? *ELT Journal, 41,* 126-131.

Sugita, Y. (2005). The effect of phrasal listening practice on Japanese college students' listening comprehension. *Kanto-Koshinetsu Association of Teachers of English (KATE) Bulletin, 19,* 11-21.

Tamai, K. (1992). "follow-up" no choukairyoku koujorni oyobosu kouka oyobi "follow-up" nouryoku to choukairyoku no kankei. *STEP Bulletin, 4,* 48-62.

Ur, P. (1984). *Teaching listening comprehension.* New York: Cambridge University Press.

Appendix A. Teaching Materials

1st Listening Recall Task (4th Grade STEP, Obunsha, 2004a, p. 27)

1. Harry wanted to play soccer with his friends yesterday, but it rained all day. So, he stayed at home and read a book.

2. Jack has two sisters, Mary and Sue. Mary is a teacher and lives in Japan. She often writes letters to Jack and tells him about her students.

2nd Listening Recall Task (Pre-2nd Grade STEP, Obunsha, 2004c, p. 28)

Last night Sara stood in front of the movie theater waiting for Neil. She waited for an hour, but he didn't come. Later Neil called to say that his car had broken down and he went to a repair shop.

3rd Listening Recall Task (Pre-2nd Grade STEP, Obunsha, 2004c, p. 29)

Robert's grandmother lives alone. Robert enjoys going to her house and talking to her. Lately she has been sick, so Robert has been visiting her house as often as possible to help her. He has also started looking for someone who can visit her when he is too busy to go.

4th Listening Recall Task (3rd Grade STEP, Obunsha, 2004b, p. 28)

On Saturday, Sally spent all day working in her garden. She picked lots of vegetables. On Sunday, she invited her friends to her house and used the vegetables from her garden to make dinner for them.

5th Listening Recall Task (Pre-2nd Grade STEP, Obunsha, 2004c, p. 48)

Mrs. Sato has two daughters. When the girls were small, she stayed at home to take care of them. But when they became teenagers, they started spending less time at home. Mrs. Sato now had a lot of time to herself, so she decided to join a volunteer group.

6th Listening Recall Task (Pre-2nd Grade STEP, Obunsha, 2004c, p. 49)

Randy's mother wanted to bake a cake, so she asked Randy to go to the store for her. He went and bought everything she needed. But on his way home, he fell down. When he got home, he saw that the eggs were broken. But his mother wasn't angry at all, so he was glad.

(2006年5月25日　受理)