

How to determine the optimal number of listening opportunities for listening comprehension tests among Japanese high school learners of English

Ken'ichi Otsuka
(Waseda University)

The purpose of this study is to suggest the optimal number of listening opportunities for listening comprehension tests used as the norm-referenced measurement among Japanese high school learner of English. This paper consists of two studies. The participants of the two studies took different listening comprehension tests respectively and they were provided three opportunities of listening for each item and the test scores were collected each time. The tests were based on STEP Pre-2nd Grade and 3rd Grade test. Test scores were analyzed by the Classical Test Theory (CTT) and the Item Response Theory (IRT). The results showed that repetition positively affected the means, reliabilities, standard error of measurement, whereas there was no statistical significance in the item discrimination power index. Furthermore, ANOVA showed that the repetitions affected positively for the learners whose listening proficiency was rated at the upper and middle level, whereas the lower level learners did not gain any advantage from repetition. These findings will be discussed in connection with the pedagogical implications of this research and the area identified for further study.

1.0 Introduction

How many listening opportunities should the test-takers be given in each item of the English listening comprehension tests if the test scores are used for a norm-referenced measurement? The global English proficiency tests such as TOEFL and TOEIC provide only one opportunity of listening for each item. On the other hand, in Japanese public high school entrance examinations, test-takers are given two listening opportunities for each item stem and question. In addition, one of the nationwide English proficiency tests, STEP tests, produced by the Society for testing English Proficiency (STEP), had provided the test takers with two opportunities of listening in its major part of listening section of 2nd Grade and 3rd Grade test until 2002. Although Step revised Pre-2nd Grade test in 2003, which newly provides item stems and questions with no repetition, they still give two opportunities of listening in the major part of 3rd Grade test.

The contention is just how many listening opportunities in a listening comprehension test is optimal for Japanese junior and senior high school students whose listening proficiency in English is normally regarded as to be elementary level. This study is intended as an investigation into the effect of the number of times repeated in English listening comprehension tests and I would like to examine whether repetition is an essential technique for succeeding in listening comprehension tests.

2.0 Theoretical background

The increased interests in how learners interact with oral input, regarding second language acquisition through listening comprehension has been examined from the point of view of several characteristics. Rubin (1994) introduced them as “1) text characteristics (variables in a listening passage, text or associated visual support); 2) interlocutor characteristics (variation in the speaker’s personal characteristics); 3) task characteristics (variations in the purpose for listening and associated responses); 4) listener characteristics (variation in the listener’s personal characteristics); 5) process characteristics (variation in the listener’s cognitive activities and in the nature of the interaction between speaker and listener)”. Furthermore, when it comes to the language testing perspective, as Bachman and Palmer (1996) pointed out, two characteristics, which are task and the test-takers’ characteristics greatly affect test performance, that is, the interactions between the test task and the situation characteristics, and the characteristics of test-takers have a strong relationship. It is obvious that more attention should be paid to these two characteristics in order to interpret a test score accurately.

The study on the task characteristics in listening comprehension has mainly been focused on how variables influence learners comprehension, such as 1) types of response (e.g., multiple choice question vs. Wh-question: Eykyn, 1992; Wu, 1998; Buck, 1991; Stansfield, 1981), 2) types of question (e.g., global question vs. local question: Shohamy and Inbar, 1991; Reed, 2000), 3) while listening activities (e.g., note taking: Hale and Courtney, 1991, 1994; Carrell et al., 2000), 4) speakers’ dialect (e.g., Fitzmaurice and Major, 2000), 5) length of a text used for dictation (Takahashi et al., 1998), 6) item format (Sherman, 1997; Yanagawa, 2003), and so on. As a result, previous studies have not paid attentions to the influence of the number of repetitions for item stem and question provided to test-takers. Although Mayer (1983) reported a positive effect of repetitions in the first language listening test in a science prose ($F(2, 42) = 5.95, p < .01, F(2, 84) = 16.33, p < .001$), little is known about the effect of repetition in a second language listening comprehension tests.

On the other hand, repetition has been widely used technique for ESL / EFL classes (Chaudron, 1988: 84-85, cited in Iimura and Ishizaki, 2001) and researches on the effect of repetition in a listening comprehension training have shown that repetitions facilitate learners’ comprehension ($F(1, 52) = 27.60, p < .0001, K-R 2I = .73, n=82, Cervantes&Gainer, 1992; t=5.08, df=38, p < .05, Taniguchi, 1999; t=.045, p < .05, Iimura and Ishizaki, 2001; F=163.245, p=.000, F=176.360, p=.000, Iimura, 2004$). Also, Yamauchi (1999) reported the positive effects on the repetitions. However, little information concerning test score interpretation has been provided from most of these previous studies because they were designed to elicit data for the use of studies in learners’ process characteristics.

As the author mentioned in 1.0, it is usual to provide two opportunities of listening for each item in Japanese listening comprehension test for novice learners such as high school students.

In order to create a valid test, it is very important that we have many items in a test because the more items that we have in a test, the more reliable that test will be (Hughes, 2003). Repetition takes at least twice as long time than the no repetition type test, and as a result, when the test was administered under the same time limitation, it might be reasonable to consider that tests with repetitions are less reliable than tests with no repetition. Reliability and validity are the essential measurement qualities and reliability is considered to be a requirement for validity (Bachman, 1990). Therefore, it is appropriate to confirm the effect of repetition in the second language listening comprehension tests.

3.0 Study 1

3.1 Method

3.1.1 Participants

Thirty-eight 12th grade Japanese co-ed senior high school students whose first language was Japanese participated in this study. They belonged to the same senior high school located in the northern part of Kanto area and their English listening proficiency was assessed to be in the intermediate to low range.

3.1.2 Materials

A set of listening tests were adapted from a listening section in STEP Pre-2nd Grade test (The Society of Testing English Proficiency, 2002) held in January 2003. The test was chosen because this test was designed for use with 11th and 12th grade students (STEP, 2004). The original test was comprised of three parts. In Part 1, there were 5 items and test-takers were to choose one of four answer options as the suitable response to a short dialogue. Each item stem was presented only once. In Part 2, there also were 5 items and test-takers were to listen to a question followed by each set of dialogues and choose the answer from four options. Each stem and question was presented twice. In Part 3, the numbers of items were 10 and test-takers were to select the suitable answers to a question related to a monologue. Stem and question were presented twice. Each item had one multiple-choice question with four alternatives and all answer options were printed in the paper and could be viewed before listening. As Figure 1 shows, an exploratory version of the test was constructed for Study 1. In this test, each item stem (and question in Part 2, and 3) was presented three times followed by 10 seconds pause for answering. In other words, all participants had three times of listening and answering opportunities respectively. They were required to write their answers on a set of answer sheets that consisted of stapled three different sheets of paper to eliminate other factors, such as priming effect which could confuse the result. One point was awarded for each correct answer.

3.1.3 Procedure

The test was conducted during regular classes in June 2004. As the students had participated in listening comprehension training in the author's class many times since the start of a school year, they were addressed in that the test scores would be considered as a part of their grade and asked to take this listening test seriously. Furthermore, no information concerning this exploratory test was provided in advance to prevent student from preparing for this listening test and becoming aware of the contents of the test prior to it.

Figure 1

A Comparison of Original Version and Exploratory Version

Part 1

Original version

<Dialogue> + 3 seconds pause + <Dialogue> + 10 seconds pause + [next item]

* Test-takers can write their answers any time before starting a next item.

Exploratory version

<Dialogue> + 10 seconds pause + <Dialogue> + 10 seconds pause + <Dialogue> +
10 seconds pause + [next item]

* Test-takers write each answer during 10 seconds pause.

Part 2

Original version

<Dialogue+Question> + 3 seconds pause + <Dialogue+Question> + 10 seconds
pause + [next item]

* Test-takers can write their answers any time before starting a next item.

Exploratory version

<Dialogue+Question> + 10 seconds pause + <Dialogue+Question> +
10 seconds pause + <Dialogue+Question> + 10 seconds pause + [next item]

* Test-takers write each answer during 10 seconds pause.

Part 3

Nearly the same as Part2.

3.2 Results and Discussions

3.2.1 Descriptive statistics

Descriptive statistics for each test score are shown in Table 1. The distributions of all the tests were proved to be reasonably normal. As this table shows, participants got the highest marks at the first time in the three different times of repetitions presented for the test-takers.

Table 1**Summary Statistics of Exploratory Version of Three Different Number of Times Repeated**

	<i>Mean</i>	<i>SD</i>	<i>Skew</i>	<i>Kurt</i>	<i>Min</i>	<i>Max</i>	<i>Full</i>	<i>Alpha</i>
Once	4.868	2.154	-0.240	-0.049	0	10	20	0.305
Twice	3.947	1.919	0.008	-1.217	1	7	20	0.264
Third	3.868	1.949	0.590	-0.087	1	9	20	0.354

Note: *N*=38. *Skew*=Skewness; *Kurt*=Kurtosis; *Min*=Minimum score; *Max*=Maximum score; *Full*=Full marks; *Alpha*=Cronbach's coefficient alpha

3.2.2 Analysis of variance

One-way analysis of variance (ANOVA) was used to examine the statistical difference between three variables. The result is presented in Table 2, which revealed that there is significance between the three groups. For further investigation of the effect of the repetitions, a Ryan's method was performed as a post-hoc test. As Table 3 shows, there were two statistical significances between Once and Third repetition(s), and Once and Twice repetition(s). And there was no significance between Twice and Third. As far as this result concerned, it appears that the participants did not gain any advantage from repetitions.

Table 2**The Result of a One-Way ANOVA for STEP Pre-2nd Grade Test Score**

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between Group	23.491	2	11.746	4.694	0.012 *
Within group	275.404	37	7.443		
Error	185.175	74	16.281		
Total	484.07	113			

**p*<.05

Table 3**The Result of a Ryan's Method Across Three Different Number of Times Repeated**

Pair	<i>r</i>	nominal level	<i>t</i> (Ryan)	<i>p</i>	<i>sig.</i>
Once - Third	3	0.017	2.756	0.007	<i>s</i>
Once - Twice	2	0.333	2.538	0.013	<i>s</i>
Twice - Third	2	0.333	0.218	0.828	<i>ns</i>

Note: *df*=74, significance level=.05, *sig*=significance

3.2.3 Item analysis and discussion

Results from item analysis implied that this test suffers from three crucial evidences for interpretation. First, the mean score of each test was quite low. Although the full mark of each test was 20, means were ranged from 3.868 to 4.868, which represent the influence of the floor effect. Second, the reliability of each test was extremely low. And third, as the number of items which had sufficient item discrimination power, which means over 0.3 in point biserial correlation coefficient, was only 7.3 out of 20 items in average (See Appendix 1). Regretfully, these evidences indicate that this exploratory test did not work well for discriminating participants appropriately. It follows from these reasons that the test contained too many difficult items for the participants and need to be revised for further study. Finally, it seems appropriate to remark that, although the result of ANOVA appeared to conclude that repetition did not positively affect the test score, useful information could not be elicited from this study.

4.0 Study 2

4.1 Method

4.1.1 Participants

One hundred and sixty-nine 12th grade Japanese co-ed senior high school students whose first language was Japanese participated in this study. They belong to the same senior high school located in the northern part of Kanto area and their English listening proficiency was assessed to be in the intermediate to low range. Thirty-eight students who took part in Study 1 were also participated in this study.

4.1.2 Materials

A set of listening test was adapted from a listening section in STEP 3rd Grade test (The Society of Testing English Proficiency, 2001, 2003) held in January 2002 and October 2003. The test was chosen because these tests were designed for use with junior high school graduate level (STEP, 2004). The 2001 version test was comprised of three parts with 20 items and the details of Part 1, 2, and 3 were the same as the test of Pre-2nd which was explained in 3.1.2. In 2003, STEP revised their Part 1 whose item had one multiple-choice question with three alternatives and all item stems and answer options were broadcasted with no repetition, that is, answer options were not printed in the paper and could not be viewed before listening, besides test-takers had only one chance of listening for item stem and answer option. As Figure 2 shows, an exploratory version of the test was constructed for Study 2. It was comprised of four parts with 18 items and each item stem (and question in Part 3, and 4) was presented three times followed by 10 seconds pause for answering, whereas the original version of the test provided different number of repetitions. In Study 2, all participants had three times of listening opportunities and answering chances respectively. They were required to write their answers in a set of answer sheet that consisted of stapled three different

sheets of paper to eliminate other factors, such as priming effect that could confuse result. One point was awarded for each correct answer.

4.1.3 Procedure

The test was conducted during a regular class in July 2004. Participants were told that the test scores would be considered as a part of their grade and asked to take this listening test seriously. Furthermore, no information concerning this test was provided in advance to prevent student from preparing for this listening test and test-wiseness. Questionnaires were administered immediately following a listening test and then, correct answers were provided. After the test, a listening test adapted from STEP 3rd grade test which was held in January 2002 was conducted under the test conditions in order to divide students into three groups, Upper, Middle, and Low Level. Based on this test scores, 43 out of 169 students whose scores ranged from 11-15 out of 18 were regarded as the upper level, 82 out of 169 were the middle level who scored from 7 to 10, and the lower level was consisted in 44 students with 3-6 points. According to the Kruskal-Wallis test, there was a statistical difference between these three groups ($H=144.732$, $p<0.0001$). Also, the result of Friedman's test conducted as a post-hoc test showed three groups were statistically different (Mann-Whitney $U=0.000$, 0.000 , 0.000 , $p<0.000$).

Figure 2

Details of a New Version Test

Part 1 (From 2003 version of Part 1) Item No. from 1 to 4.

Part 2 (From 2001 version of Part 1) Item No. from 5 to 8.

Part 3 (From 2003 version of Part 2) Item No. from 9 to 13.

Exploratory version

<Dialogue> + 10 seconds pause + <Dialogue> + 10 seconds pause + <Dialogue> +
10 seconds pause + [next item]

* Test-takers write each answer during 10 seconds pause.

Part 4 (From 2003 version of Part 3) Item No. from 14 to 18.

Exploratory version

<Monologue> + 10 seconds pause + < Monologue > + 10 seconds pause + < Monologue > +
10 seconds pause + [next item]

* Test-takers write each answer during 10 seconds pause.

4.2 Results and Discussions

4.2.1 Descriptive statistics and CTT analysis

Descriptive statistics for each test score are shown in Table 4. The distributions of all the tests were proved to be reasonably normal. As this table shows, participants got the highest marks at

the third in the three different times of repetitions presented for the test-takers.

Table 4

Summary Statistics of Exploratory Version for Study 2

	<i>Mean</i>	<i>SD</i>	<i>Skew</i>	<i>Kurt</i>	<i>Min</i>	<i>Max</i>	<i>Full</i>	<i>Alpha</i>
Once	8.751	2.456	0.405	-0.227	3	16	20	0.348
Twice	9.485	2.795	-0.044	-0.140	1	17	20	0.503
Third	9.568	2.874	0.186	-0.337	4	18	20	0.533

Note: $N=169$. *Skew*=Skewness; *Kurt*=Kurtosis; *Min*=Minimum score; *Max*=Maximum score;

Full=Full marks; *Alpha*=Cronbach's coefficient alpha

Cronbach's reliability coefficient alpha was calculated by the computer program "Test Data Analysis Program" (TDAP) created by Ohtomo, Nakamura, and Kiyama (2002). Whereas it appeared that all the reliability coefficients were fairly low, it is reasonable to suggest that repetitions positively affected the reliability. According to Nakamura and Ohtomo (2002), the benchmark of the Cronbach's coefficient alpha is 0.8. To make each test's reliability higher than 0.8, that is, the level of reliability desired, TDAP showed the following number of items which should be added to the new test; 135 items for the Once version, 71 items for the Twice version, and 63 items for the Third version. In terms of feasibility of doing this kind of listening comprehension test, it is apparently hard to add such a large number of items under the time limitation. Since the original version of STEP 3rd test, which had 20 items with twice of repetitions, was conducted approximately within 20 minutes, the test used for Study 2 could be revisable to the no repetition version with roughly 40 items under the same time limitation. Using Ohtomo's formula (Ohtomo, 1996), it was found that 0.69 of reliability coefficient could be gained by doubling the number of test items of no repetition version, that is, the test used for the Study 2. In other words, if the test with no repetition version has 36 items, the reliability of the test could be much higher than versions with twice or third repetitions. As to the reliability coefficient, it is reasonable to suppose that doubling the number of item makes the test more reliable than offering twice or third repetitions.

4.2.2 Analysis of variance

To examine the statistical difference between three variables, One-way analysis of variance (ANOVA) was used. The result is presented in Table 5, which revealed that there is a statistical difference between the three groups. To further investigate the difference between the three groups, a Ryan's method test was performed as a post-hoc test. As Table 6 shows, there were statistical significances between Third repetitions and Once, and Twice and Third repetitions. And there was no significance between Third and Twice group. As far as this result concerned, it appears that the participants did gain advantage from repetitions.

Table 5***The Result of a One-Way ANOVA for STEP 3rd Grade Test Score***

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>
Between Group	68.276	2	34.138	17.709	0.0001**
Within group	3087.519	168	18.378		
Error	647.724	336	1.928		
Total	3087.519	506			

p*<.01**Table 6**The Result of a Ryan's Method Across Three Different Number of Times Repeated***

Pair	<i>r</i>	nominal level	<i>t (Ryan)</i>	<i>p</i>	<i>sig.</i>
Third - Once	3	0.016	5.406	0.000	<i>s</i>
Third - Twice	2	0.333	0.548	0.583	<i>ns</i>
Twice - Third	2	0.333	4.858	0.000	<i>s</i>

Note: *df*=336, significance level=0.05, *sig*=significance**4.2.3 Analysis of item discrimination power index**

For the test administrators, one of the concerns is to know whether items are activating for discriminating the test-takers' proficiency appropriately. To confirm this, interpretation of the item discrimination power is crucial. In Table 7, the number of items which had an appropriate item discrimination power index (IDPI), that is, over 0.30 in point biserial correlation coefficient, the means on IDPI of each item and its Standard deviation are shown (See Appendix 2 for further information). Means and *SDs* was calculated by the scores after being transformed by Fisher's *z*'-transformation formula. To examine the effect on the number of times repeated, Friedman's test was conducted. According to the result of Friedman's test, the number of repetitions did not make any statistical difference between the three groups (*Friedman* $\chi^2=3.00$, *df*=2, *p*<.05 *ns*). As far as this result is concerned, it appears that the repetitions did not aid discriminating test-takers proficiency.

Table 7***Number of Items which Has Appropriate IDPI and Means of IDPI***

	<i>N</i>	<i>N's of over 0.3 IDPI</i>	mean	<i>SD</i>
Once	18	10	0.297	0.057
Twice	18	13	0.340	0.070
Third	18	12	0.349	0.100

Note: IDPI= item discrimination power index; *N's* = numbers; Means and *SDs* of IDPI were calculated after all IDPIs were transformed by Fisher's *z*'-transformation formula.

4.2.4 Analysis of Standard Error of Measurement in PROX

Another important factor for test score interpretation is the standard error of measurement (*SEM*). To examine the relationship between the effect of repetitions and *SEM* on each participant, one parameter Rasch model was used and all responses of the test were transformed by TDAP. Friedman's test shows that there is a statistical significance between three versions in .01 level (*Friedman* $\chi^2=22.357$, $df=2$, $p<0.0001$). For further investigation, Scheffe test was conducted as a post-hoc test. The result showed that there found significances between Once and Third version in .01 level, and Second and Third version in .05 level.

4.2.5 Brief summery of discussions

For summery, it was observed in the proceeding section that reliability formula, means of test score, and *SEM* index marked the highest when the item stems, questions, and answer options were repeated third, while there was no significance in *IDPI*.

4.2.6 Two-way analysis of variance

The further investigation was conducted to reveal the effect of interactions between participants' listening comprehension proficiency and the number of times repeated. Table 8 shows the difference of means and *SDs* between three different proficiency levels and in Table 9, the result of the Two-way ANOVA is shown followed by Ryan's method test conducted as a post-hoc (Table 10 and 11). As Two-way ANOVA shows, the interaction between proficiency and the number of times repeated was found and then the simple main effect was examined. Table 12-18 shows the results. According to the result of the simple main effect test, it is concluded that the participants in upper level got the greater effect on repetition of twice rather than that of third, in contrast, middle level participants got the great effect at the third repetitions. Moreover, lower level participants did not get any positive effects on repetitions.

Table 8

Difference of Mean and SD between Three Different Proficiency Levels

	<i>n</i>	Once		Twice		Third	
		Mean	<i>SD</i>	Mean	<i>SD</i>	Mean	<i>SD</i>
Upper level	43	11.070	2.214	13.023	1.248	12.488	2.117
Middle level	82	8.671	1.828	9.512	1.085	9.610	1.924
Lower level	44	6.636	1.553	5.977	1.323	6.636	1.872

Table 9***The Result of a Two-Way ANOVA for STEP 3rd Grade Test Score******(Proficiency / Number of repetitions)***

Source	SS	df	MS	F	p
Proficiency	2581.701	2	1290.850	235.457	0.000****
Error	910.064	166	5.482		
N of repetitions	58.266	2	29.133	16.912	0.000****
P - N	89.260	4	22.315	12.954	0.000****
Error	571.899	332	1.753		

**** $p < .001$ **Table 10*****The Result of a Ryan's Method for Means of Proficiency***

Pair	r	nominal level	t (Ryan)	p	sig.
upper - lower	3	0.017	19.929	0.000	s
upper - middle	2	0.033	11.510	0.000	s
middle - lower	2	0.033	11.272	0.000	s

Note: $df=166$, significance level=0.05, sig=significance**Table 11*****The Result of a Ryan's Method for Means of Number of times repeated***

Pair	r	nominal level	t (Ryan)	p	sig.
Third - Once	3	0.017	5.504	0.000	s
Third - Twice	2	0.033	0.518	0.605	ns
Twice - Once	2	0.033	4.986	0.000	s

Note: $df=332$, significance level=0.05, sig=significance**Table 12*****The Result of a Simple Main Effect Analysis Across******Proficiency and the Number of Times Repeated***

Effect	SS	df	MS	F	p
A(B1)	507.907	2	253.953	85.339	0.0000***
A(B2)	1280.017	2	640.009	215.069	0.0000***
A(B3)	883.037	2	441.518	148.368	0.0000***
Error		498	2.976		

(table continues)

Table 12 (continued)

Effect	SS	df	MS	F	p
B(A1)	105.102	2	52.551	30.507	0.0000***
B(A2)	27.49	2	13.745	7.979	0.0004***
B(A3)	14.933	2	7.467	4.335	0.0139*
Error		332	1.723		

Note: A=proficiency, B= number of times

repeated; A1= upper, A2= middle, A3=

* $p < .05$, *** $p < .001$

lower; B1=Once , B2= Twice, B3= Third

Table 13***The Result of a Ryan's Method Means on Proficiency / Once***

Pair	r	nominal level	t (Ryan)	p	sig.
upper - lower	3	0.017	11.985	0.000	s
upper - middle	2	0.033	7.386	0.000	s
middle - lower	2	0.033	6.311	0.000	s

Note: df=498, significance level=0.05, sig=significance

Table 14***The Result of a Ryan's Method Means on Proficiency / Twice***

Pair	r	nominal level	t (Ryan)	p	sig.
upper - lower	3	0.017	19.048	0.000	s
upper - middle	2	0.033	10.810	0.000	s
middle - lower	2	0.033	10.965	0.000	s

Note: df=498, significance level=0.05, sig=significance

Table 15***The Result of a Ryan's Method Means on Proficiency / Third***

Pair	r	nominal level	t (Ryan)	p	sig.
upper - lower	3	0.017	15.82	0.000	s
upper - middle	2	0.033	8.863	0.000	s
middle - lower	2	0.033	9.224	0.000	s

Note: df=498, significance level=0.05, sig=significance

Table 16***The Result of a Ryan's Method Means on Number of Times Repeated / Upper Level***

<i>Pair</i>	<i>r</i>	<i>nominal level</i>	<i>t (Ryan)</i>	<i>p</i>	<i>sig.</i>
Twice - Once	3	0.017	6.901	0.000	<i>s</i>
Twice - Third	2	0.033	1.890	0.600	<i>ns</i>
Third - Once	2	0.033	5.012	0.000	<i>s</i>

Note: *df*=332, significance level=0.05, *sig*=significance

Table 17***The Result of a Ryan's Method Means on Number of Times Repeated / Middle Level***

<i>Pair</i>	<i>r</i>	<i>nominal level</i>	<i>t (Ryan)</i>	<i>p</i>	<i>sig.</i>
Third - Once	3	0.017	4.581	0.000	<i>s</i>
Third - Twice	2	0.033	0.476	0.634	<i>ns</i>
Twice - Once	2	0.033	4.105	0.000	<i>s</i>

Note: *df*=332, significance level=0.05, *sig*=significance

Table 18***The Result of a Ryan's Method Means on Number of Times Repeated / Lower Level***

<i>Pair</i>	<i>r</i>	<i>nominal level</i>	<i>t (Ryan)</i>	<i>p</i>	<i>sig.</i>
Third - Once	3	0.017	2.355	0.019	<i>ns</i>
Third - Twice	2	0.033	0.000	1.000	<i>ns</i>
Twice - Once	2	0.033	2.355	0.190	<i>ns</i>

Note: *df*=332, significance level=0.05, *sig*=significance

4.2.7 Result from the questionnaire

To examine test-takers' reaction to this test, short questionnaires including 4 multiple choice and one open ended questions were designed (Appendix 3). This questionnaire was designed as a pilot study for the further study. The questionnaires were conducted right after the listening comprehension test, that is, before the correct answers of each item were provided to test-takers and furthermore, test-takers were told that the answers of the questionnaire would never affect their grade to eliminate factors which can confuse participants' answer of questionnaires.

Since the questionnaire was conducted as a pilot study, the result collected from this questionnaire was too small and provided only a little information, however, the result shows that most of the test-takers tend to prefer to listen to item options and questions twice or third, on the other hand, there were few test-takers who want to listen only once, or fourth or more. There were no interaction between the test-takers' proficiency level and their preference to the number of times

repeated (Appendix 4). From an open-ended question, the following tendencies were found. 1) the more test-takers have listening opportunities, the more anxiety decrease, 2) over fourth repetitions decrease test-takers concentration.

4.2.8 Summary of discussions

So far the author has showed the results and discussions of Study 2, to sum up major characteristics briefly as follows. 1) Reliability: Third > Twice > Once, Doubled Once > Third > Twice > Once. 2) Means: Third > Once, Third = Twice, Twice > Once. Means; Upper level: Twice > Once, Twice = Third, Third > Once. Middle level: Third > Once, Twice = Thirds, Twice > Once, Lower level: Thirds = Twice = Once. 3) Item Discrimination Power Index: Thirds = Twice = Once. 4) Standard Error of Measurement: Third > Once, Third > Twice.

As far as these results concerned, in the author's understanding, repetition cannot be an essential method of testing listening comprehension proficiency for Japanese senior high school level participants because what needs to be emphasized at language testing, particularly in a language proficiency testing paradigm whose scores are usually used as a norm-referenced measurement, is that whether each test items are discriminating between test-takers language proficiency. The result of this study clearly showed that repetition of item stem or question did not aid IDPIs whereas the author recognizes the importance of the positive effects on means and reliabilities. Furthermore, although repetition affects the interactions between means and proficiency levels complexly, it does not negatively influence discrimination.

The result of a short questionnaire indicated that test-taker's reaction should be investigated in further study because it appears that test-taker's anxiety and concentration have much to do with test administration and interpretation of test score.

5.0 Conclusion

The current study has examined to what extent repetitions of item stem and questions in a English listening test influence test-takers' score and test score interpretation from the point of view of test administrators. Participants took listening comprehension tests which were revised to provide three opportunities of listening in item stems and questions, followed by a questionnaire. The mean score of each test showed that repetitions appeared to facilitate test-takers' listening comprehension at the upper and middle level, whereas low level participants did not gain any effect from repetitions. Above all, other analysis indicated that while means, reliability, and *SEM* are positively affected by the repetitions, listening test could be undertaken without any repetition when the numbers of items were sufficient. It is my hope that the findings of the current research help decisions makers draw more valid inference from test scores.

References

- Bachman, L.F. 1990. *Fundamental Consideration in Language Testing*. Oxford; Oxford University Press.
- Bachman, L.F. and Palmer, A. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.
- Buck, G. 1991: The test of listening comprehension: An introspective Study. *Language Testing*, 8, 1, 67-91.
- Carrell, P., Dunkel, P., and Mollan, P. 2000. The effects of notetaking, stimulus length, and topic on the listening component of TOEFL 2000. *Paper presented at the American Association of Applied Linguistics Annual Meeting*, Vancouver, Canada. March 11-14, 2000.
- Cervantes, R and G, Gainer. 1992: The Effect of Syntactic Simplification and Repetition on Listening Comprehension. *TESOL Quarterly*, 26, 767-770.
- Chaudran, C. 1988. *Second Language Classrooms*. Cambridge: Cambridge University Press.
- Eykyn, L. B. 1992. *The Effects of Listening Guides on the Comprehension of Authentic Texts by Novice Learners of French as a Second Language*. Diss., Univ. of South Carolina.
- Fitzmaurice, S. and Major, R. 2000. The effect of different dialects on listening comprehension. *Paper presented at the American Association of Applied Linguistics Annual Meeting*, Vancouver, Canada. March 11-14, 2000.
- Hale, G.A. and Courtney, R. 1991. Note-taking on listening comprehension on the Test of English as a Foreign Language. *TOEFL Research Reports*, 34. Princeton, NJ: Educational Testing Service.
- and Courtney, R. 1994. The effect of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing*, 11, 1, 29-47.
- Hughes, A. 2003: *Testing for Language Teachers Second Edition*. Cambridge: Cambridge University Press.
- Imura, H., and Ishizaki, T. 2001. The Effect of Pauses on Listening Comprehension – Repetition vs. Pauses -, *Tsukuba Eigo Kyouiku (Tsukuba Journal of English Education)*, 22, 117-124.
- Imura, H. 2004. Risuningu no kaisuu to risuningu purosusu no kankei. *Paper presented at the 30th Conference of Zenkoku Eigo Kyouiku Gakkai (The Society of English Language Education)*, Nagano, Japan. 2004.
- In'nami, Y. In press. The effects of test-takers' reactions to task types and test anxiety on listening test performance, *Unpublished paper given personally by Yo In'nami in September, 2004*.
- Mayer, R.E. 1983. Can you repeat that? Qualitative effects of repetition and advanced organizers on learning from science prose. *Journal of Educational Psychology*, 75, 40-49.
- Nakamura, Y. and Ohtomo, K. 2002: *Tesuto de gengo nouryoku wa hakareru ka – gengo tesuto deita bunseki nyuumon*. Tokyo: Kiriara Shoten.

- Ohtomo, Nakamura, and Akiyama. 2002: Test Data Analysis Program (TDAP) Ver.2.0 [Windows version]. Tokyo: Kirihara Shoten. (Attached CD-ROM to Nakamura, Y. and Otomo, K. 2002).
- Ohtomo, K. 1996. *Koumoku Outou Riron Nyuumon*. Tokyo: Taishukan Shoten.
- Reed, J. 2000: Effects of interactive input in assessing listening. *Paper presented at Teachers of English to Speakers of Other Language, Inc. 34th Annual Convention*. Vancouver, Canada. March 14-18, 2000.
- Rubin, J. 1994. A Review of Second Language Listening Comprehension Research. *The Modern Language Journal*, 78, 199-221.
- Sherman, J. 1997. The effect of question preview in listening comprehension tests. *Language Testing*, 14, 2, 185-213.
- Shohamy, E. and Inbar, O. 1991. Validation of listening comprehension tests: the effects of text and question type. *Language Testing*, 8, 1, 23-40.
- Stanfield, C.W. 1981. Dictation as a measure of Spanish language proficiency. *IRAL*, 19, 4.
- STEP. 2004. <http://www.eiken.or.jp/english/evaluate/index.html>
- Takahashi, H., Shiina, K. and Takefuta, Y. 1988. Dai 1 bu Hiaringu ni kansuru kenkyu I. Keishiki no ninshiki – kyouzai, oyobi jizen jyohou no teijihou -. *Gengo Koudou no Kenkyu* 4, 124-138. University of Chiba.
- Takei, A. 2002. *Eigo risuningu ron: kiku chikara to shidou wo kagakusuru*. Japan. Kagensha.
- Taniguchi, Y. 1999. Risuningu kaisuu ga risuningu rikai ni oyobosu kouka. *STEP BULLETIN*, 12, 26- 35.
- The Society of Testing English Proficiency. 2002a. *Monbu kagaku shou nintei jitsuyou eigo kentei Pre 2nd Grade test form*. 2002-3.
 ----- 2002b. *Monbu kagaku shou nintei jitsuyou eigo kentei 3rd Grade test form*. 2002-3.
 ----- 2003. *Monbu kagaku shou nintei jitsuyou eigo kentei 3rd Grade test form*. 2003-2.
- Wu, Y. 1998. What do tests of listening comprehension test? – A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15, 1, 21-44.
- Yamauchi Y. 1999: Eigo Risuningu ni okeru kurikaesi ga rikai ni oyobosu eikyou
 - Purosesu Jyushi no risuningu shidou no yuukousei -.
Tokyo Kokusai Daigaku Ronsou Shougakubu Hen, 59, 89-106.
- Yanagawa, K. 2003. Relative Difficulty of Three Multiple Choice Listening Comprehension Item Formats for Japanese High School Learners of English. *The Bulletin of the Graduate School of Education of Waseda University Separate Volume* No.10-2, 1-14.

Appendix 1

Item Discrimination Power of each item in Study 1

(Point biserial correlation coefficient)

	Once	Twice	Thirds
Item No.1	0.227	0.453	0.511
Item No.2	0.449	0.385	0.247
Item No.3	0.241	0.305	0.589
Item No.4	0.281	0.367	0.592
Item No.5	0.342	0.418	0.080
Item No.6	0.369	0.296	0.278
Item No.7	0.151	0.296	0.281
Item No.8	-0.012	0.367	0.346
Item No.9	0.326	0.297	0.243
Item No.10	0.199	0.278	0.466
Item No.11	0.301	0.313	0.106
Item No.12	0.203	-0.022	-0.094
Item No.13	0.241	0.155	0.463
Item No.14	0.235	0.214	0.264
Item No.15	0.465	0.173	0.076
Item No.16	-0.019	0.099	0.029
Item No.17	0.379	0.099	0.370
Item No.18	0.443	0.191	0.220
Item No.19	0.204	0.200	0.095
Item No.20	0.173	0.188	0.243

Appendix 2

Item Discrimination Power of each item in Study 2

(Point biserial correlation coefficient)

	Once	Twice	Thirds
Item No.1	0.241	0.310	0.336
Item No.2	0.327	0.272	0.267
Item No.3	0.330	0.359	0.339
Item No.4	0.224	0.312	0.341
Item No.5	0.327	0.467	0.518
Item No.6	0.347	0.32	0.317
Item No.7	0.332	0.342	0.387
Item No.8	0.261	0.337	0.257
Item No.9	0.245	0.348	0.283
Item No.10	0.218	0.340	0.323
Item No.11	0.312	0.342	0.456
Item No.12	0.321	0.274	0.357
Item No.13	0.325	0.303	0.228
Item No.14	0.303	0.277	0.278
Item No.15	0.240	0.188	0.138
Item No.16	0.188	0.273	0.356
Item No.17	0.264	0.391	0.373
Item No.18	0.372	0.411	0.438

Appendix 3

Questionnaire for test-takers.

-
- Q1. How many times would you like to listen to in Part 1? Once - Twice - Thirds - Fourth or more
- Q2. How many times would you like to listen to in Part 2? Once - Twice - Thirds - Fourth or more
- Q3. How many times would you like to listen to in Part 3? Once - Twice - Thirds - Fourth or more
- Q4. How many times would you like to listen to in Part 4? Once - Twice - Thirds - Fourth or more
- Q5. Write any thing you felt toward this listening test.

(Keyword; number of times repeated, concentration, confidence, anxiety, test score, grade, etc.)

Note: Original questionnaire was written in Japanese.

Appendix 4

The result of a Questionnaire

"How many times would you like to listen to in a listening test?"

Part 1

	Once	Twice	Third	Fourth or more
Upper	13%	54%	30%	3%
Middle	3%	50%	41%	6%
Lower	2%	54%	42%	2%

Part 2

	Once	Twice	Third	Fourth or more
Upper	3%	47%	50%	0%
Middle	1%	46%	49%	4%
Lower	2%	47%	44%	7%

Part 3

	Once	Twice	Third	Fourth or more
Upper	3%	54%	40%	3%
Middle	1%	41%	52%	6%
Lower	2%	49%	40%	9%

Part 4

	Once	Twice	Third	Fourth or more
Upper	3%	40%	50%	7%
Middle	0%	47%	43%	10%
Lower	2%	42%	47%	9%

Note: Upper = 43 out of 169 participants rated at the Upper level;

Middle = 82 out of 169 participants rated at the Middle level;

Lower = 44 out of 169 participants rated at the Lower level