Research Paper

Factors Contributing to Holistic Listening of Kyoto University Students A Preliminary Study

Masayasu Aotani

Abstract

Different aspects of holistic listening comprehension, as found in TOEFL iBT, of Japanese-speaking Kyoto University students were studied with a battery of 15 tests. There were strong correlations among the scores of listening comprehension, reading comprehension, and listening cloze tests. On the other hand, more microscopic and local aural skills of phoneme distinction and word recognition were much more weakly correlated. Grammar and syntax knowledge was also shown to contribute little to holistic listening comprehension for the students. These findings were corroborated by the results of an exploratory principal component analysis. Multiple-regression analyses showed that about 70 percent of the variance in holistic listening comprehension was accounted for by reading comprehension of listening scripts and listening cloze. The fact that these tests are much better predictors of holistic listening ability than listening comprehension of short conversations seems to indicate the students' heavy reliance on the skills of integrating information, following the general logical flow, and grasping the main ideas of the text. Though more local skills of sound and word recognition as well as grammatical and syntactic analysis/processing did not explain much variance in the students' performance, their significance may be masked by the fairly uniform grammatical/syntactic abilities of the students as well as the uniform weaknesses in sound distinction and aural word recognition.

【Keyword】 listening comprehension, reading comprehension, Rasch analysis, multiple regression, TOEFL

1. Introduction

Some teachers of English at Kyoto University have long known that students perform better on a long listening comprehension test, like the listening section of TOEFL iBT, than a short listening comprehension test, like the 10-second conversations in TOEFL PBT. When their average TOEFL iBT section scores were compared with the average of all Japanese examinees and all examinees irrespective of the nationality, it appeared that Kyoto University students' long listening comprehension was on a par with that of all the examinees worldwide. However, an unpublished study conducted in 2009 (Aotani) showed that their average short listening comprehension score was significantly lower than that of all the examinees. The students' mean score was 14.5 out of 30 compared with 21.0 for all the examinees, and the difference between the means was significant with $p < .0005$.

Table 1. ***Average TOEFL iBT Scores of All Japanese Examinees, Kyoto University Students, and All Examinees***

	Reading	Listening	Speaking	Writing	Total
Japanese	16	16	15	18	65
Kyoto	23	19	15	22	78
All	19	20	19	20	78

Note. Kyoto University students' averages were computed based on the scores voluntarily submitted in response to our request. The averages of Japanese examinees and all examinees were provided by the Educational Testing Service (ETS).

In this research project, 112 Japanese-speaking Kyoto University students' holistic listening comprehension was studied with a battery of 15 tests to create a profile of their proficiency at various listening subskills. In particular, comparisons were made between long/holistic listening and short listening.

It is with listening that humans start their linguistic activities as a newborn or even before birth (Brown, 1987; Jalongo, 2010; Robinshaw, 2007). However, this fundamental linguistic skill is often regarded as the least researched of the four language skills (Chand, 2007), the other three being speaking, reading, and writing. This is clearly illustrated by the much smaller number of publications on listening than on reading, the other receptive skill of reading. One approach to listening comprehension research is to try to decompose listening skill into its component subskills. One can study both underlying psycholinguistic parameters such as working memory capacity, as well as the traditionally recognized subskills of listening, including word recognition and syntactic knowledge. This project's goals were to study (1) the contributions made by the subskills of the latter type, (2) what different types of tests contribute to accounting for holistic listening, and (3) the difference (s) between short listening and holistic listening.

2. Materials and methods

2.1. Participants

The participants in this study were 112 Kyoto University students, mostly freshmen, who were taking the author's English classes offered in the spring semester of 2010. Their mother tongue was Japanese. Actual numbers of examinees differed from one test to another due mainly to absences, and I included in this study only those 112 students who did not miss any examinations. These students had all received at least six years of English education in junior and senior high schools and passed a highly competitive entrance examination of Kyoto University, whose English section consists exclusively of translations from English to Japanese and vice versa.

2.2. Instruments

I used a battery of 15 tests to measure the students' listening comprehension and its subskills, as well as other linguistic abilities potentially contributing to listening comprehension. The characteristics of the tests are summarized in the table below (Table 2).

Table 2. ***Battery of 15 Tests in the Present Investigation***

Test Name (Variable Name)	Short Description
Long Listening Test (LLT)	Lectures and conversations lasting for 2 to 5 minutes followed by 5 to 6 multiple-choice questions
Short Listening Test (SLT)	A short conversation with two turns followed by a multiple-choice question
Dictation Test (DTN)	Three short and three long sentences with a total of 77 words
Aural Word Recognition Test (AWR)	Dictation of 50 difficult words
Phoneme Distinction Test: Consonant Pairs (PHD)	Five pairs of words for each of the following contrasts are presented: [b, v], [f, h], [l, r] (at the beginning and in the middle of the word), and [s, th] (k = 25)
Listening Cloze Test (LCT)	Deletion of every tenth word in three listening passages (k = 80)
Reading Cloze Test (RCT)	Deletion of every tenth word in three reading passages (k = 80)
Vocabulary Size Test (VST) (Beglar, 2010; Nation & Beglar, 2007)	Multiple-choice vocabulary items; 3,000 level through 8,000 frequency level
Productive Vocabulary Levels Test (PVT) (Laufer & Nation, 1999)	Controlled productive ability of words at the 2,000 and 3,000 word frequency levels (k = 18)
Grammatical Error Detection Test (GED)	Finding a segment containing a grammatical error from four underlined segments in each sentence
GGT (See *Note* below.)	GFT + GGF
Structure: Gap Filling Test (GFT)	Selecting the correct expression to fill a gap in each sentence (This test is for nonnative speakers.)
GRE Gap Filling Test (GGF)	Selecting the correct expression to fill a gap or two in each sentence (This test is for native speakers.)
Reading Comprehension of Listening Scripts Test (RCL)	Reading the scripts for long listening tests and answering 5 to 6 multiple-choice questions
Metacognitive Awareness Listening Questionnaire (MALQ, MALQJ) (Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006)	Test of the knowledge of and the ability to monitor, control, and integrate cognitive processes important for listening. (MALQJ is a Japanese version.)

Note. Each test has a three- or four-letter acronym, which serves both as the test name and variable name. k = the number of items. As GFT and GGF are both gap-filling tests of a similar nature, these tests were later combined via Rasch analysis to produce a new better-performing test/variable GGT.

2.3. Methods

The result of each test was subjected to Rasch analysis (1980), and the person ability measures instead of the raw scores were used for correlation studies, multiple regression, and principal component analysis. One exception was the dictation test (DTN) which was not conducive to Rasch type item-wise analysis as even the concept of "items" was not clear there. Students were encouraged to guess what they were not able to hear clearly in this test, often resulting in different numbers of words for each sentence.

Scores that were more than 3 standard deviations away from the mean were regarded as outliers. Those

scores were replaced by

[the score that corresponds to SD = +3 (or – 3)] + 1 (or – 1)

prior to further investigations. There were no multivariate outliers as judged by Mahalanobis distance.

2.3.1. Validation of Instruments

Misfitting items detected by the initial Rasch analysis were removed, and the Rasch procedure was repeated until both person and item parameters were satisfactory and unidimensionality conditions are generally satisfied. I first checked if there was a reverse polarity problem as evidenced by a negative Pt-measure correlation. Any item with a negative Pt-measure correlation coefficient was excluded unconditionally from the rest of the analysis, and Rasch analysis was run again with the new set of items after the deletion of such a misfitting item. The second stage was checking Infit MNSQ and Outfit MNSQ item by item to pick out misfitting items which are candidates for deletion. Wright suggested $(1-2/\sqrt{N}, 1+2/\sqrt{N})$ as the acceptable range for Infit MNSQ and $(1-6/\sqrt{N}, 1+6/\sqrt{N})$ as the acceptable range for Outfit MNSQ (Smith, 2001), where N is the sample size or the number of students who took the test. Beglar, in personal communication, suggested that $(M-2SD, M+2SD)$ be used as the acceptable range both for Infit MNSQ and Outfit MNSQ; where M is the mean of Infit and Outfit MNSQ respectively, and likewise for SD. Because I used this as the first step towards eliminating misfitting items, I felt it was a good idea to pick out as many candidate items for later deletion as possible. Therefore, I used the interval

$$(\mathrm{Max}\{1-2/\sqrt{N}, M-2SD\}, \mathrm{Min}\{1+2/\sqrt{N}, M+2SD\})$$

as the acceptable range. Once the candidates for deletion were picked, each item was checked in detail as follows.

From the list of "MOST MISFITTING RESPONSE STRINGS" provided by Winsteps (Linacre, 2007), persons with unexpected responses, up to 5% of the sample size, were removed and the fit statistics were rechecked for the item. If it is still misfitting, the item was tagged for deletion. Otherwise, it became a candidate for retention. Some apparently misfitting items did not have an entry in the list of most misfitting response strings, indicating that the responses fit the model rather well. These items were retained despite their fit parameters. These procedures are summarized in Table 3.

Table 3. ***Criteria and Actions for the Investigation of Item Fit***

Criterion	Critical Value/Range	Action
Item Polarity	Point-measure Correlation > 0	Item is removed if the value < 0.
Infit MNSQ	All figures should be strictly in the range bounded from below by $\mathrm{Max}\{1-2/\sqrt{N}, \mathrm{Mean}-2\mathrm{SD's}\}$ and from above by $\mathrm{Min}\{1+2/\sqrt{N}, \mathrm{Mean}+2\mathrm{SD's}\}$.	For items outside the range, the "MOST MISFITTING RESPONSE STRING" criterion below is applied.
Outfit MNSQ	All figures should be strictly in the range bounded from below by $\mathrm{Max}\{1-6/\sqrt{N}, \mathrm{Mean}-2\mathrm{SD's}\}$ and from above by $\mathrm{Min}\{1+6/\sqrt{N}, \mathrm{Mean}+2\mathrm{SD's}\}$.	For items outside the range, the "MOST MISFITTING RESPONSE STRING" criterion below is applied.
Most Misfitting Response String	Responses expected by the Rasch model	Persons with unexpected responses (up to 5% of the sample size *N*) are removed and item fit statistics are rechecked.

Note. *N* is the sample size, and Mean is the mean for Infit MNSQ and Outfit MNSQ respectively. These criteria were applied in steps, starting from the top of the table. The "Most Misfitting Response String" criterion applies only to items that failed to meet either the Infit MNSQ or Outfit MNSQ criterion. If the item still misfits after the persons with unexpected responses have been removed, it will be tagged for deletion and subjected to further investigation.

Items thus selected as potential candidates for deletion were actually removed only after a close examination of their effects on student and item separation and reliability, item-person map, the variances explained, and unidimensionality. The item-person map was examined for the range of measurement, overall matching between item difficulty and person ability, and potential gaps in the measurement scale. Unidimensionality conditions are described in Table 4.

Table 4. ***Criteria for Investigating Unidimensionality***

Criterion	Critical Value/Range
Item Reliability	Reliablity ≥ .95
Item Separation	Separation ≥ 3.0
Variance Explained	To be determined by the "Targeting-Explained Variance" graph for dichotomous data in the Winsteps Manual (Linacre, n.d.-c)
Empirical-Modeled Comparison	Reasonable agreement between the observed and modeled values (Linacre, n.d.-c) : 10% difference is acceptable (Linacre, n.d.-c).
First Contrast's Eigenvalue	Smaller than 2 is ideal. Smaller than 3 is acceptable. (Linacre, n.d.-b)
Ratio of the variances explained by items and by the first contrast	Above 4 is desirable (Linacre, n.d.-a).
Residual Plot	The points should not be separated into two distinct groups. (visual inspection)

Note. The eigenvalue of the first contrast can be as large as 2 due exclusively to random error (Raîche, n.d.). Empirical-Modeled Comparison is a measure of unidimensionality and/or model fit as modeled variances mean "variance components expected for these data if they exactly fit the Rasch model, i.e., the variance that would be explained if the data accorded with the Rasch definition of unidimensionality" (Linacre, n.d.-c). I did not apply each criterion in a step-by-step fashion. Instead, a unidimensionality judgment was made in an integrative and holistic manner, taking all these factors into account.

Due to space limitations, I will only briefly sketch the validation procedure for the Rasch model for LLT (the Long Listening Test).

The Long Listening Test (LLT) had the following items with misfitting parameters.

Table 5. ***Misfitting Items***

	Infit MNSQ	Outfit MNSQ
Acceptable Range	(.83, 1.17)	(.68, 1.28)
Misfitting Items	9, 12, 15	7, 9, 15, 36

After the removal of up to 7 persons (146*.05 = 7.3; i.e. 5% of the sample size) with most misfitting response strings, I obtained the results summarized in Table 6.

Table 6. ***Fit Statistics After the Deletion Of Persons With Unexpected Responses***

Item	7	9	12	15	36
Measure (before, after)	-3.48, -6.13	.06, -3.74	.15, .03	.57, .56	-3.66, -6.15
Persons Removed (Sample Size)	27, 94, 118, 121 (142)	40, 75, 80, 87 (142)	18, 75, 85, 87, 121 (141)	25, 27, 41 (143)	43, 66, 110 (143)
$(M-2SD, M+2SD)_i$	(.82, 1.18)	(.82, 1.18)	(.82, 1.18)	(.84, 1.16)	(.82, 1.18)
(1-2/sqrtN, 1+2/sqrtN)	(.832, 1.168)	(.832, 1.168)	(.832,)	(.833, 1.167)	(.833, 1.167)
Infit Range	(.832, 1.168)	(.832, 1.168)	(.834, 1.168)	(.84, 1.16)	(.833, 1.167)
Infit MNSQ	MINIMUM	1.15	1.10	1.21	MINIMUM
$(M-2SD, M+2SD)_o$	(.70, 1.26)	(.72, 1.24)	(.73, 1.25)	(.70,1.26)	(.73, 1.25)
(1-6/sqrtN, 1+6/sqrtN)	(.496, 1.504)	(.496, 1.504)	(.495, 1.505)	(.498, 1.502)	(.498, 1.502)
Outfit Range	(.70, 1.26)	(.72, 1.24)	(.73, 1.25)	(.70, 1.26)	(.73, 1.25)
Outfit MNSQ	MINIMUM	1.14	1.10	1.20	MINIMUM
Deletion/Retention	Deletion	Retention	Retention	Deletion	Deletion

Note. Measure (before, after) means that the figure on the left is the item difficulty measure before deletion, and the figure on the right is the same measure after deletion. M is the mean, SD is the standard deviation, $(M-2SD, M+2SD)_i$ and $(M-2SD, M+2SD)_o$ refer to the Infit and Outfit MNSQ respectively, and sqrtN signifies the positive square root of the sample size N. Infit and Outfit Range define the acceptable values of Infit and Outfit MNSQ.

Items 7, 15, and 36 were tagged for deletion as a result of this series of examinations. However, upon a closer examination of item and person parameters as well as unidimensionality conditions, I decided to delete only Item 15. This led to the parameters presented in Table 7 below.

Table 7. ***Parameters after the Deletion of Item 15***

TEST	Long Listening Test (LLT)			
DELETIONS	15			
	Mean Measure	SD	Separation	Reliability
STUDENTS	.37	.81	1.79	.76
ITEMS	0	1.17	5.06	.96
TOTAL VAR.	24.3			
	Eigenvalue		%	
ITEM VAR.	8.6		16.6	
CONTRAST	2.3		4.4	
ITEM/CONTRAST	16.6/4.4 = 3.77			

Note. "Deletions" means the deleted item (s). "TOTAL VAR.", "ITEM VAR.", and "CONTRAST" mean "Raw variance explained by measures", "Raw Variance explained by items", and "Unexplained variance in 1st contrast". "ITEM/CONTRAST" means the ratio of "Raw Variance explained by items" to "Unexplained variance in 1st contrast".

Furthermore, variances explained by this model were as shown in Table 8.

Table 8. ***Variances Explained for LLT (standardized residual variance in eigenvalue units)***

	Empirical		Modeled
	Eigenvalue	% of Variance	% of Variance
Total raw variance in observations	51.5	100.0%	100.0%
Raw variance explained by measures	12.5	24.3%	24.4%
Raw variance explained by persons	3.9	7.7%	7.7%
Raw Variance explained by items	8.6	16.6%	16.7%
Raw unexplained variance (total)	39.0	75.7%	75.6%
Unexplained variance in 1st contrast	2.3	4.4%	
Unexplained variance in 2nd contrast	2.1	4.1%	
Unexplained variance in 3rd contrast	2.0	3.8%	
Unexplained variance in 4th contrast	1.9	3.7%	
Unexplained variance in 5th contrast	1.8	3.5%	

It is rare that all the parameters lie in the desired ranges and all the unidimensionality conditions are completely satisfied. There is always an element of subjective judgment, but with this understanding, the Long Listening Test satisfied the validation criteria when Item 15 was removed. Figures 1 and 2 show the item-person map for LLT and loadings plot for the first contrast.

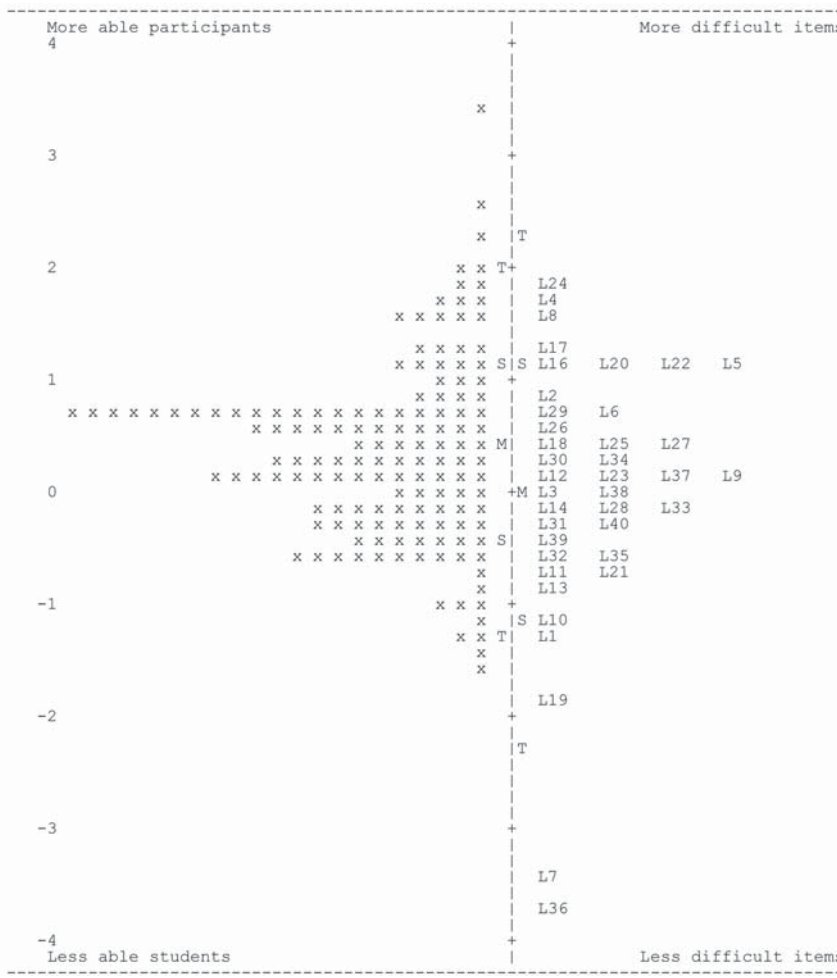

Figure 1. Item-Person Map for the Long Listening Test

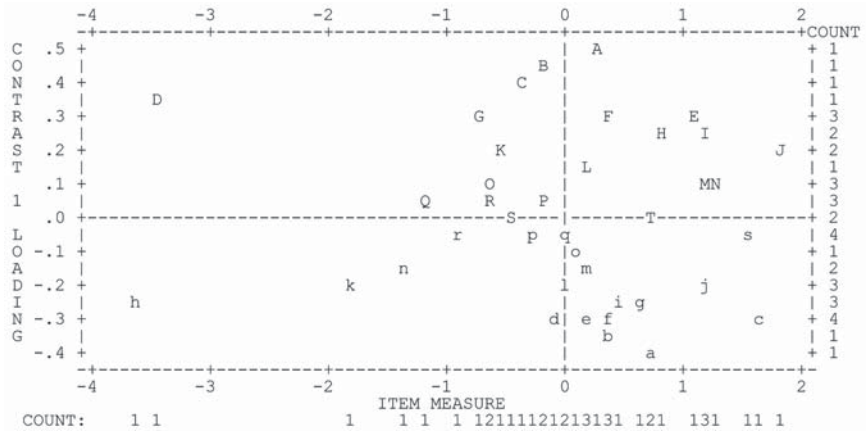

Figure 2. Standardized residual contrast 1 plot

Similar validation procedures were applied to the remaining 13 tests/variables.

3. Results

3.1 Correlation studies

Pearson correlations among the ability measures for all the tests after applying Rasch analysis are given in Table 9. The ability measure for Test XYZ is denoted by M-XYZ. Note that GFT (the Gap Filling Test) and GGF (the GRE Gap Filling Test) were combined to produce a new measure denoted by GGT as explained in the note of Table 2. This is a simple sum of the items in GFT and GGF, and hence, has a total of 65 items. Also note that most of the 66 significance levels were at $p < .0005$ except for M-LLT vs. M-PHD (.002), M-AWR vs. M-GGT (.002), M-PHD vs. M-RCL (.017), M-GGT vs. M-PHD (.006), M-GED vs. M-GGT (.004), M-PHD vs. M-VST (.001), and M-PHD vs. M-RCT (.008). Therefore, the probability of finding at least one bivariate correlation erroneously flagged as significant in this table is at most .0673, where .0673 would be achieved only when we use $p = .0005$ for the 59 pairs for which $p < .0005$. We should note here that SPSS shows .000 for all p-values smaller than .0005, and all these cases are reported only as $p < .0005$ in Table 9. Actual p-values are significantly smaller than .0005 for many of the 59 pairs mentioned above. As a result, no such measure as Bonferroni adjustment was necessary despite the large number of entries in the correlation table.

It was immediately obvious that M-MALQ and M-MALQJ were not significantly correlated with any other variable but with themselves. While the low but significant positive correlation between M-MALQ and M-MALQJ may be an indicator of some level of consistency and construct validity of this instrument as a standalone measure of metacognitive awareness about listening, these instruments were not useful for the purposes of explaining listening comprehension of Kyoto University students. Hence both M-MALQ and M-MALQJ were dropped from further analyses.

This analysis mainly concerns M-LLT (long listening comprehension), whose correlation coefficients

with other variables ranged from .295 with M-PHD to .756 with M-RCL (Figures 3 and 4).

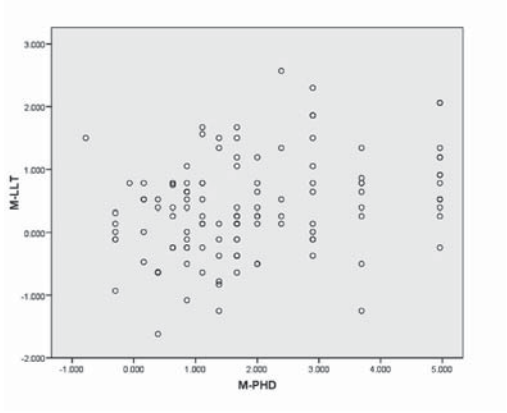

Figure 3. M-LLT vs. M-PHD Plot

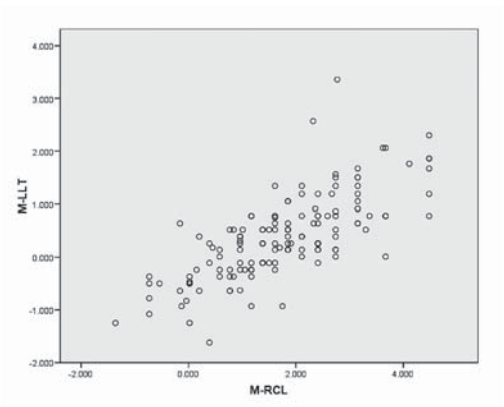

Figure 4. M-LLT vs. M-RCL Plot

The person ability measure M-LLT is based on listening comprehension tests with the listening section consisting of a 3 to 5 minutes' lecture or conversation, which is then followed by 5 to 6 questions. This is qualitatively different from M-SLT where the students answer one question after listening to a 10-second exchange between two interlocutors. When the columns labeled 1 and 2 in Table 9 are compared, one can see there is a tendency for SLT (short listening) to be more strongly correlated with genuinely aural tests such as DTN (dictation), AWR (aural word recognition), and PHD (phoneme distinction). On the other hand, LLT (long listening) appears to be more strongly correlated with tests that require context-based judgment and more holistic understanding of the text. Typical examples include LCT (listening cloze), RCT (reading cloze), GGT (see *Note* of Table 2), and RCL (reading comprehension). Particularly notable is the fact that RCT, GGT, and RCL require reading and no listening. In order to quantify and confirm/disconfirm these observations, which are nothing but visual inspections at this point, I computed Steiger's z (1980) using FZT.exe (Garbin, n.d.). Steiger's z was used rather than the traditional favorite Hotteling's t because it has been pointed out that the former is a better measure for this purpose (Meng, Rosenthal, & Rubin, 1992). Under certain circumstances Hotteling's t may overestimate the t-value, leading to a Type I error. The results are summarized in Table 10.

Table 9. *Steiger's Z to Compare Correlations*

Variable	1. M-LLT	2. M-SLT	Steiger's Z	Steiger's Z (N = 200)
1. M-LLT	——	.621**	——	——
2. M-SLT	.621**	——	——	——
3. Z-DTN	.507**	.576**	-1.026	-1.380
4. M-AWR	.353**	.489**	-1.838	-2.471*
5. M-PHD	.295**	.409**	-1.481	-1.991*
6. M-LCT	.705**	.664**	0.740	0.994

7. M-RCT	.666**	.584**	1.347	1.811
8. M-VST	.555**	.571**	-0.243	-0.327
9. M-PVT	.457**	.518**	-0.864	-1.162
10. M-GED	.603**	.556**	0.728	0.978
11. M-GGT	.457**	.379**	1.051	1.413
12. M-RCL	.756**	.645**	2.080*	2.797**

Note. The symbol "**" in the columns labeled M-LLT and M-SLT means the correlation is significant at $p < .01$. The symbol "*" and "**" in the columns labeled Steiger's Z and Steiger's Z (N = 200) mean the correlation is significant at $p < .05$ and $p < .01$ respectively. For $p < .05$, the two-tailed critical value of Z is 1.96, and it is 2.58 for $p < .01$. The column labeled Steiger's Z (N = 200) is a hypothetical column where the Z-values are presented for a hypothetical case of N = 200 instead of the actual 112. This was done in order to examine the effect of the sample size, which is rather small in this case. The sample size N enters the formula for Steiger's Z as $\sqrt{N-3}$ in the numerator.

The results seem to indicate that the difference for reading (RCL) is indeed significant and the differences for AWR and PHD are close to being significant at $p < .05$. This is also supported by the hypothetical Z-values for N = 200.

Another notable fact is that M-LLT (long listening) is correlated more strongly with M-RCL (reading), M-LCT (listening Cloze), and M-RCT (reading Cloze) than with M-SLT (short listening). This is despite the fact that LLT and SLT both require aural processing, among other abilities, and have multiple choice questions. In order to further investigate the relations among different variables, I tried multiple regression studies to explain the variances in LLT and SLT in terms of the other test scores and principal component analyses to probe into latent constructs.

3.2 Multiple regression studies

Multiple regression analyses were conducted mainly to investigate how much of the variance in listening comprehension was accounted for by the other tests. This offers one way to understand what subskills are important for holistic listening comprehension (LLT, M-LLT), which is the focus of this investigation. The dependent variable I focused on was M-LLT.

Few things in this world are truly linear, and linear modeling is often very inadequate in that sense. Nevertheless, Box was quite right when he remarked in his book: "Essentially, all models are wrong, but some are useful (Box & Norman, 1987, p. 424)." In addition, multiple linear regression is one of the most commonly employed tools in the studies of language skills and subskills (Chiappe, Siegel, & Gottardo, 2002; Chiappe, Siegel, & Wade-Woolley, 2002; Feyten, 1991; Geva, Yaghoub-Zadeh, & Schuster, 2000; Geva & Yaghoub Zadeh, 2006; Grabe, 2009; Lesaux, Koda, Siegel, & Shanahan, 2006; Limbos & Geva, 2001; Mecartty, 2000; Vandergrift, 2006; Wade-Woolley & Siegel, 1997), and the abundance of literature makes it easy to compare the results with other related studies if multiple regression is used. It is for these reasons that I also used multiple regression.

Though we should never expect a relationship to be perfectly describable by a linear model, some

datasets violate the conditions for the use of this tool so severely that they render multiple regression useless. Despite the comfort we can take in the prevalent and generally successful use of this technique in language acquisition studies, it is necessary to make sure the most basic conditions for multiple regression analysis are indeed satisfied (Osborne & Waters, n.d.). Because the distributions of my variables are generally symmetric, if not perfectly normal, judging from the Item-Person maps obtained at the end of Rasch analysis, the most basic first condition appears to be met. The rest of the basic conditions will be checked after a model is selected.

As shown in Table 9, both the Reading Comprehension of Listening Scripts Test (RCL, M-RCL) and the Listening Cloze Test (LCT, M-LCT) are highly correlated with M-LLT (holistic listening comprehension of long conversations and lectures). Guided by this observation and the fact that backward elimination starting with all variables other than M-LLT as independent variables led to the set (M-LCT, M-RCL, M-VST, M-PVT), four distinct models with different independent variables were examined. Additionally, one model which was the best fit for M-SLT was included as Model 5. Only the data for the 112 participants with no missing values were included.

Model 1: All variables other than M-LLT as independent variables

Model 2: M-RCL and M-LCT as independent variables with a constant term

Model 3: M-RCL and M-LCT as independent variables without a constant term (Note: This was done after observing that the constant term was not significant at $p < .05$ in Model 2.)

Model 4: M-RCL, M-LCT, M-VST, and M-PVT as independent variables (Note: This was tried after observing that backward elimination starting with all variables led to these four.)

Model 5: M-RCL, M-LCT, and M-PHD as independent variables

Table 10. ***Full Correlation Matrix***

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. M-LLT	——	.621***	.507***	.353***	.295**	.705***	.666***	.555***	.457***	.603***	.457***	.756***	0.051	0.023
2. M-SLT	.621***	——	.576***	.489***	.409***	.664***	.584***	.571***	.518***	.556***	.379***	.645***	0.076	0.009
3. Z-DTN	.507***	.576***	——	.478***	.386***	.659***	.539***	.425***	.470***	.564***	.400***	.507***	0.019	-0.012
4. M-AWR	.353***	.489***	.478***	——	.417***	.610***	.435***	.426***	.506***	.483***	.284**	.321***	0.127	0.073
5. M-PHD	.295**	.409***	.386***	.417***	——	.375***	.248**	.297**	.335***	.265***	.249**	.223*	-0.030	-0.037
6. M-LCT	.705***	.664***	.659***	.610***	.375***	——	.658***	.570***	.544***	.616***	.408***	.583***	0.076	0.000
7. M-RCT	.666***	.584***	.539***	.435***	.248**	.658***	——	.573***	.611***	.748***	.542***	.719***	-0.082	-0.116
8. M-VST	.555***	.571***	.425***	.426***	.297**	.570***	.573***	——	.655***	.538***	.574***	.581***	0.159	-0.017
9. M-PVT	.457***	.518***	.470***	.506***	.335***	.544***	.611***	.655***	——	.691***	.637***	.527***	0.089	-0.003
10. M-GED	.603***	.556***	.564***	.483***	.265***	.616***	.748***	.538***	.691***	——	.689**	.669***	0.045	-0.005
11. M-GGT	.457***	.379***	.400***	.284**	.249**	.408***	.542***	.574***	.637***	.689**	——	.544***	0.066	-0.015
12. M-RCL	.756***	.645***	.507***	.321***	.223*	.583***	.719***	.581***	.527***	.669***	.544***	——	-0.014	0.021
13. M-MALQ	.051	.076	.019	.127	-.030	.076	-.082	.159	.089	.045	.066	-.014	——	.340**
14. M-MALQJ	.023	.009	-.012	.073	-.037	.000	-.116	-.017	-.003	-.005	-.015	.021	.340**	——
Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14

* $p < .05$ ** $p < .01$ *** $p < .0005$ Listwise $N = 112$

The specifications for the above five models are summarized in Table 11 below.

Table 11. ***Five Multiple Regression Models***

Model	Dependent Variable	Independent Variables	Constant
1	M-LLT	M-SLT, M-VST, M-PVT, M-GED, M-GGT, M-RCL, M-LCT, M-RCT, Z-DTN, M-AWR, M-PHD	Yes
2	M-LLT	M-LCT, M-RCL	Yes
3	M-LLT	M-LCT, M-RCL	No
4	M-LLT	M-LCT, M-RCL, M-VST, M-PVT	Yes
5	M-SLT	M-LCT, M-RCL, M-PHD	Yes

The results are summarized in Table 12 below.

Table 12. ***Results of Five Multiple Regression Analyses***

Model Summary

Model	R	R^2	Adjusted R^2	SE
1	.829	.687	.653	.460
2	.810	.655	.650	.460
3	.844	.711[a]	.707[a]	.459
4	.811	.657	.646	.464
5	.756	.572	.561	.539

ANOVA

Model	SS	df	MS		Sig.
1	Regression	47.804	11	4.346	.000
	Residual	21.816	103	.212	
	Total	69.620	114		
2	Regression	48.694	2	24.347	.000
	Residual	25.606	121	.212	
	Total	74.300	123		
3	Regression	63.272	2	31.636	.000
	Residual	25.656	122	.210	
	Total	88.928	124		
4	Regression	48.716	4	12.179	.000
	Residual	25.392	118	.215	
	Total	74.108	122		
5	Regression	44.717	3	14.906	.000
	Residual	33.428	115	.291	
	Total	78.145	118		

Coefficients

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	*SE*	Beta	*t*	*p*
1	(Constant)	.066	.240		.277	.783
	M-SLT	.044	.083	.046	.525	.601
	M-VST	.074	.106	.059	.692	.491
	M-PVT	-.080	.089	-.081	-.900	.370
	M-GED	.002	.094	.002	.019	.985

	M-GGT	.003	.103	.003	.032	.974
	M-RCL	.286	.055	.473	5.171	.000
	M-LCT	.372	.088	.407	4.234	.000
	M-RCT	.075	.104	.071	.723	.471
	Z-DTN	-.009	.059	-.011	-.147	.884
	M-AWR	-.094	.070	-.099	-1.338	.184
	M-PHD	.023	.035	.043	.672	.503
2	(Constant)	.057	.116		.488	.627
	M-RCL	.335	.060	.364	5.575	.000
	M-LCT	.333	.040	.543	8.316	.000
3	M-LCT	.312	.036	.446	8.595	.000
	M-RCL	.350	.020	.887	17.111	.000
4	(Constant)	.017	.131		.128	.898
	M-LCT	.333	.066	.361	5.033	.000
	M-RCL	.325	.044	.530	7.355	.000
	M-PVT	.087	.093	.070	.929	.355
	M-VST	-.047	.074	-.048	-.642	.522
5	(Constant)	-.309	.160		-1.936	.055
	M-PHD	.097	.036	.175	2.673	.009
	M-LCT	.362	.074	.380	4.908	.000
	M-RCL	.246	.047	.387	5.217	.000

[a]It is inappropriate to use R^2 and adjusted R^2 for the purpose of comparing a model without a constant term with one that includes a constant term. A better measure in this case appears to be the pair-wise correlation between the predicted values and the observed values, and that is what I used (Eisenhauer, 2003; Hocking, 1996, p. 178). See the main text.

The variable M-LLT is best described by a linear combination of M-LCT and M-RCL which successfully accounted for 67.5% (square of the pair-wise correlation coefficient between observed and predicted values: See the note under Table 12.) of the variance in M-LLT. We have

$$\text{M-LLT} = (0.312)\ \text{M-LCT} + (0.350)\ \text{M-RCL}.$$

Parenthetically, the correlation between the values predicted by Model 2 and Model 3 was .998. The choice of Model 2 is due to its simplicity.

I next checked for violations of regression assumptions for this model (Model 2). The scatter plot of studentized residuals against unstandardized predicted values showed no sign of curvilinearity (Norušis, 2008, pp. 262-263) (Figure 5). Furthermore, the scatter plot of standardized residuals against standardized predicted values appeared rectangular, indicating homoscedasticity (Statistics Solutions, n.d.) (Figure 6). Therefore, the most basic regression assumptions were satisfied for this model.

For M-SLT, I arrived at

$$\text{M-SLT} = (0.362)\ \text{M-LCT} + (0.246)\ \text{M-RCL} + (0.097)\ \text{M-PHD} - 0.309.$$

It is worth noting that both the Simple View of Reading ($R = D \times C$) (Gough & Tunmer, 1986; Hoover & Gough, 1990) and the Component Model of Reading ($R = D \times C + S$) (Aaron, 1997; Joshi, 1999; Joshi & Aaron, 2000) have a product of listening comprehension C and decoding skill D, as well as an additional

processing speed S for the latter model, in their formulas for reading comprehension R. What is interesting is the fact that the proponents of the Component View of Reading mention in one of their papers that addition (R = D + C) and multiplication (R = D × C) did not always produce very different predictions (Joshi & Aaron, 2000, p. 90). This suggests the possibility of a similar product model for listening comprehension, which can explain as much variance as our linear regression model.

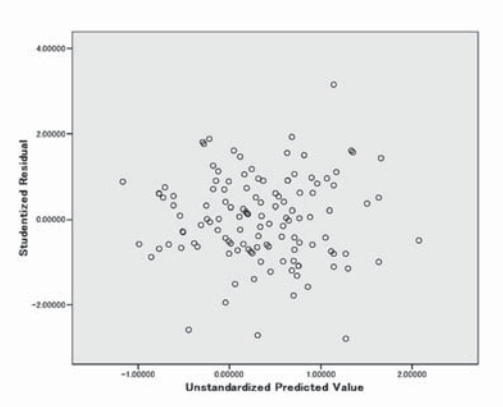

Figure 5. Studentized residuals vs. unstandardized predicted values

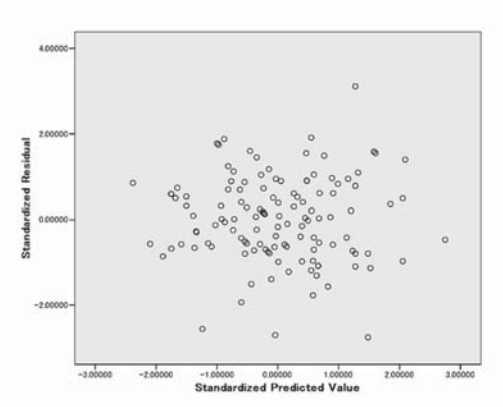

Figure 6. Standardized residuals vs. standardized predicted values

3.3. Exploratory principal component analysis

Principal component analyses were conducted with SPSS (2007). The variables included in the initial analysis were M-LLT, M-SLT, Z-DTN, M-AWR, M-PHD, M-LCT, M-RCT, M-VST, M-PVT, M-GED, M-GGT, M-RCL, M-MALQ, and M-MALQJ. In order to ensure sampling adequacy, Keiser-Meyer-Olkin Measure of Sampling Adequacy (KMO measure) as well as Keiser-Meyer-Olkin Measures for Individual Variables (KMO measures for individual variables) were checked first. The initial KMO measure was .902 which is "marvelous" according to Kaiser (1974), while the KMO measures for individual variables were .498 and .475 for M-MALQ and M-MALQJ respectively, and ranged from .866 to .944 for other variables. Values above .65 are considered reasonably large by Norušis (2008, p. 395). With all these variables included, there were three distinct components. However, one component was defined by M-MALQ and M-MALQJ, which were correlated with virtually no other variables than with each other in previous analyses. This provided another reason for removing MALQ/M-MALQ and MALQJ/M-MALQJ from further analysis. After dropping the MALQ variables, I obtained the following rotated component matrix (Table 13).

Table 13. ***Rotated Component Matrix***

	Component	
	1	2
M-LLT	0.717	
M-SLT	0.552	0.589
Z-DTN	0.465	0.603
M-AWR		0.758
M-PHD		0.783
M-LCT	0.585	0.626
M-RCT	0.811	
M-VST	0.709	
M-PVT	0.710	
M-GED	0.818	
M-GGT	0.770	
M-RCL	0.823	

Note. Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Eigenvalue > 1.0. Loadings smaller than .35 have been dropped.

I will ignore M-LLT and M-SLT for now and focus only on the other 10 variables. First, note that highest loadings on Component 1 are for M-RCL (reading comprehension), M-GED (finding grammatical errors), and M-RCT (reading cloze) in that order. This component seems to be associated with the availability of written text and its processing at different levels. Secondly, Component 2 exhibits high loadings for M-PHD (phoneme distinction) and M-AWR (aural word recognition) and seems to be associated with analytic listening at different local levels. Both Z-DTN (dictation) and M-LCT (listening cloze) require listening as well as guessing/inferring the meanings of difficult parts from the rest of the sentence/text, and this inference requires the same skills that are needed for processing written English. In fact, students had the entire text with gaps available for LCT and were able to guess the missing words from what they had been able to hear and write down in DTN. These may be the reasons why Z-DTN and M-LCT load on both components.

As for M-LLT and M-SLT, most M-LLT questions and all the answer choices were presented in written format, either on paper or on the screen, and all M-SLT answer choices were written on paper. This partly explains their loadings on Component 1. Recalling that Component 2 is basically about different types of "genuine" listening such as sound and word recognition, it is quite notable that M-LLT does not load strongly on Component 2. The cutoff for entry into the Component 2 column is .35, and the loading for M-LLT on Component 2 was .339, about the same level as M-PVT. I have previously noted that M-RCL and M-LCT best explain the variance in M-LLT (Table 12), with M-RCL being a slightly stronger predictor. Exploratory principal component analysis seems to corroborate the view that M-LLT is aligned more closely with reading comprehension and other types of processing of written texts than with the more genuine listening skills required for AWR (Aural Word Recognition). Here, the emphasis in "processing of written texts" should probably be placed on "processing (ability) " and not on "written text". Further discussions will follow.

4. Discussion

Qualitative differences between LLT and SLT are reflected on the correlation coefficients (Table 9), multiple regression studies, and the results of exploratory principal component analyses (Table 13).

In the correlation study, the correlation between M-LLT and M-SLT was smaller than the correlations for pairs (M-LLT, M-LCT), (M-LLT, M-RCT), (M-LLT, M-RCL), (M-SLT, M-LCT), and (M-SLT, M-RCL). Furthermore, Steiger's z indicated that the difference between the correlation coefficients of M-RCL with M-LLT and M-SLT is significant at $p = .038$. The significance levels were at $p = .066$ and $p = .14$ respectively for M-AWR and M-PHD. Partial correlations between M-LCT and M-RCL were insignificant when controlled for M-LLT and .252 when controlled for M-SLT. The independent variables M-RCL and M-LCT are both strong predictors of M-LLT and M-SLT, with M-RCL being the leading predictor for M-LLT and with M-LCT probably the more important for M-SLT. However, the association/correlation between M-LCT and M-RCL is more closely aligned with M-LLT than with M-SLT as is reflected in the above mentioned partial correlation coefficients.

Multiple regression studies showed that M-RCL (reading) is the leading predictor of M-LLT (long listening), and M-LCT (listening cloze) is the leading predictor of M-SLT (short listening), though they both have large coefficients in the linear formula for M-LLT and M-SLT. Again, partial correlations between M-LCT and M-RCL were insignificant when controlled for M-LLT and .252 at $p < .005$ when controlled for M-SLT. The correlation between M-LCT and M-RCL seems to be mostly a result of the sharing of the same latent variable (s) with M-LLT rather than with M-SLT. In fact, this is exactly what the principal component analyses indicated. With multiple regression approximation, I was able to explain about 70% of the variance in M-LLT. There has not been a study exactly comparable to this one, but similar attempts to explain variances in L2 reading or listening seem to leave 30 to 50% of the variances unexplained (Bernhardt, 2005; Feyten, 1991; Mecartty, 2000; Vandergrift, 2006). We may be seeing a ceiling effect for the forceful endeavor to "linearize" a relationship that is really not linear. On the other hand, the remaining variance, so far unaccounted for, may indicate the existence of and/or the necessity for other factors and parameters such as those of psycholinguistic nature. But, that would be beyond the scope of the current research.

One question that may arise is about the use of a long listening type script to obtain M-RCL. This could indeed induce bias in favor of M-LLT when the correlation coefficients were computed with M-LLT and M-SLT. However, the purpose of using listening scripts for the reading test in this study was to compare reading comprehension with holistic listening comprehension, which required similar types of text. Therefore, the use of listening scripts for reading comprehension was a part of the research design, and finding whether it leads to a "bias" in the form of a higher correlation coefficient with M-LLT is a part of my research objective. In fact, in an as-yet unpublished data set, 35 students were given both a short listening test (SLT) and a reading test with short listening type scripts and questions. The correlation between the raw scores was only .607 with $p < .0005$ as compared with .756 with $p < .0005$ for M-RCL and M-LLT. This seems consistent with the fact that listening-specific skills such as sound recognition are more important for

SLT, whereas higher level understanding of the text is more important for LLT.

The first stage of input processing is capturing and recognizing incoming bits and pieces of language, which is followed by the second stage where grammatical, syntactic, and contextual knowledge are employed to comprehend the message. It is the difficulty associated with this first stage that is a larger factor for M-SLT than M-LLT. Comparing listening and reading comprehension of similar texts is a good way to measure the relative importance of the two stages as the success of the first stage is typically automatic for reading. Word recognition is a very good example of this. Due to the inter-word spacing convention of writing as well as the ease of visual recognition as opposed to aural recognition, word recognition in reading is always nearly automatic.

Note that item-wise performance for the 35 students is unknown for this unpublished data set, and it was not possible to apply the Rasch procedure here. For the sake of comparison, for LLT and RCL, the raw scores before the Rasch analysis correlated at .733 with $p < .0005$ as compared with .756 at $p < .0005$ for M-LLT and M-RCL obtained via the Rasch procedure. Therefore, it is not unreasonable to expect that the above difference between .756 and .607 indeed reflects a real difference.

Principal component analysis generated two major components. Component 1, which accounts for the largest amount of variance in the sample, seems to be associated with performances when a written text is available, whereas Component 2 seems to represent genuine listening. Hence, when a test requires further processing after capturing the sounds, the corresponding variable should load on both components. Consistent with this general interpretation, M-SLT, which requires a balanced combination of recognition of sounds and words and further linguistic processing, loads equally on the two components as expected. On this account, it is remarkable that M-LLT, a listening test, loads exclusively on Component 1, the non-listening component. In other words, one achieves a high score on LLT not because of a superior "genuine" listening skill, which consists of perceiving sounds and recognizing words, but because of other abilities necessary during the linguistic processing for and of meaning. This fact seems to be consistent with the fact, mentioned at the beginning of this paper, that Kyoto University students do better on longer and more holistic listening tasks than their short listening counterparts.

This study has clearly shown some of the reasons why Kyoto University students do better on long listening tests than short listening tests. Long listening tests require the ability to integrate and organize the information scattered over the text, often as bits and pieces, in a systematic fashion to construct a holistic understanding of the entire text. This is possible only if one has the ability to follow the flow of the argument or exchanges through the long text. This goes beyond the simple recognition of sounds and words of English. Sentence-level comprehension is an essential infra-structure of this skill, but that alone is not sufficient to arrive at the understanding of the main message of the text or parts thereof. On the other hand, if the construction of meaning is successful as an ongoing process during listening, some missed words and the meaning of incomprehensible sentences could be guessed from the context and the meaning of the whole text. In other words, failings and shortcomings of bottom-up processing can be compensated for by top-down

strategies in long listening tests.

One version of the threshold hypothesis about reading (Alderson, 1984) asserts that one's reading skill in L1 transfers to L2 reading comprehension once that person has acquired sufficient basic proficiency in the second language. I do not necessarily agree that there is a noticeable threshold, but I beleive there is indeed a transfer of reading skills across languages based on his own first-hand experience and numerous observations of past students. Two different types of "transfer" seem to be at work for the students.

The first is between L1 reading and L2 reading. Kyoto University students' L1 reading comprehension is generally very good, as their high school grades and various test scores indicate. While their basic linguistic proficiency in English may be insufficient as a whole, their reading comprehension in English is respectable as judged by the above-average TOEFL reading section scores (Table 1) for example. I believe this is attributable to the transfer of their L1 reading comprehension to L2. While bottom-up processing, especially its front-end skills such as recognition of sounds and orthography, is rather strongly language specific, the ability to read and comprehend a long text shares much in common across human languages as indicated by the above-mentioned threshold hypothesis for reading comprehension.

The second is a "transfer" between reading and listening. Developmentally speaking, in L1, one learns to listen first and applies that skill and knowledge to reading in grade school. However, for the students in this study, most of whom went through a conventional Japanese-style English education with a strong emphasis on reading in secondary school, the order is different. By the time they enter college and start their training to improve listening, they have a fairly solid reading skill. The order of skills development is thus reversed for them. In the language of psycholinguistics and neuroscience, such a phenomenon may be referred to as "sharing" or "borrowing" instead of "transferring". That is why I wrote "transfer" in quotation marks rather than just transfer without them. Nevertheless, the focus is not on what actually happens in the brain, but on the fact that acquired and existing skills, including those developed in other processing modes such as reading, seem to be useful for listening comprehension.

To be sure, there is the Unitary Process View (Danks, 1980; Perfetti, 1985; Royer, 1985; Sanders, 1977; Sinatra, 1990; Sticht, Beck, Hank, Kleiman, & James, 1974) and there also is the Dual Process View (Brown, 1994; Lund, 1991; Maeng, 2006; Mecartty, 2000; Murphy, 1996; Park, 2004; Thompson, 1995). The Unitary Process View asserts that listening and reading are essentially the same except for the mode of input, while the Dual Process View claims there are significant differences between listening and reading. The fact is that evidence is now mounting in support of the Dual Process View. For example, reading is a better mode for paying attention to detail, but listeners can grasp and recall the main ideas of a text better (Lund, 1991). However, scholars also have a general and strong agreement that listening comprehension and reading comprehension are still quite alike in many ways requiring very similar abilities. Therefore, the Dual Process View is not in contradiction with the "transfer" or "sharing" hypotheses supported by many of my observations.

As for the dichotomy of bottom-up and top-down processing for listening comprehension, both seem to be able to benefit from reading comprehension skill. I already mentioned that the front end of the bottom-up process is strongly language-dependent. However, once the words and sentences are captured and the meaning of each part is deciphered, the remainder of the meaning formation is quite similar to that in reading. On the other hand, top-down processes are more language independent by nature. It seems reasonable to regard the use of context and world knowledge, for example, as a fairly universal process for different languages. Because each sentence lasts only for about ten seconds, and also because the total number of sentences is often three or four for SLT, any failure at the front end, such as failing to capture a certain sound and recognize some word, may render listening comprehension impossible. Compensating for a missing piece is far easier for LLT where one has the luxury of 3 or more minutes of time and the larger volume of text proportional to the longer time span.

My investigation confirmed, in a more precise quantifiable manner, what many researchers already "knew." Kyoto University students score well on long listening tests, taking full advantage of their ability to read English. This casts some serious doubts about the potential shortcomings of the currently fashionable long format listening tests. At least for the students in this study and others with a similar proficiency profile, a listening test should include both a long listening comprehension section and a sound- and word-recognition section, as I did in this study.

One obvious implication of these findings for learners of English at Japan's leading universities is that they should focus more on the first stage of the bottom-up listening process such as sound perception and word recognition. Too much emphasis and reliance on higher level comprehension and top-down processing may be a major weakness of many college students in Japan. In other words, a better balance between bottom-up and top-down skills is desirable. Many students seem to rely heavily on holistic skills by necessity due to their weak sound perception and word recognition abilities.

Finally, reiterating in a nutshell what has already been explained, I found many reasons to believe that the transfer and sharing of skills actually occur, and are very useful in processing English in different modes. Teachers of English should be fully aware of this and interpret the students' test scores accordingly. In particular, as mentioned already, my findings indicate that what is known as holistic listening comprehension tests are more like a reading comprehension test than a test of all-around listening skill. At the same time, Japanese learners of English should feel encouraged by this confirmation that L1 skills and skills of reading in English are both useful for listening comprehension in English.

Acknowledgment

I would like to sincerely thank Dr. David Beglar of Temple University Japan for his guidance, understanding, patience, and most importantly, his constant encouragement and friendship. His mentorship was indispensable in acquiring well-rounded knowledge and skills as a TESOL researcher. It is only with his continued professional and personal support that I was able to complete this long and demanding project.

References

(1) Aaron, P. G. (1997). The impending demise of the discrepancy formula. *Review of Educational Research, 67*, 461-502.

(2) Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1-24). London: Longman.

(3) Aotani, M. (2009). Proficiency profile of Kyoto University students. Unpublished research. The International Center, Kyoto University.

(4) Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27* (1), 101-118.

(5) Bernhardt, E. B. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25* (-1), 133-150.

(6) Box, G. E. P., & Norman, R. D. (1987). *Empirical model-building and response surfaces*: John Wiley & Sons, Inc.

(7) Brown, H. D. (1994). *Teaching by principles: An introductive approach to language pedagogy*. Englewood Cliffs, NJ: Prentice-Hall Regents.

(8) Chand, R. K. (2007). Same size doesn't fit all: Insights from research on listening skills at the University of the South Pacific (USP). *International Review of Research in Open and Distance Learning 8* (3), 1-22.

(9) Chiappe, P., Siegel, L. S., & Gottardo, A. (2002). Reading-related skills of kindergartners from diverse linguistic backgrounds. *Applied Psycholinguistics, 23* (1), 95-116.

(10) Chiappe, P., Siegel, L. S., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading, 6* (4), 369 - 400.

(11) Danks, J. (1980). Comprehension in listening and reading: Same or different? In J. Danks & K. Pezdek (Eds.), *Teaching English as a second or foreign language* (pp. 271-294). San Diego, CA: Academic.

(12) Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics, 25* (3), 76-80.

(13) Feyten, C. M. (1991). The power of listening ability: An overlooked dimension in language acquisition. *Modern Language Journal, 75* (2), 173-180.

(14) Garbin, C. P. (n.d.). FZT Computator Retrieved 09/27, 2010, from http://psych.unl.edu/psycrs/statpage/FZT.exe

(15) Geva, E., Yaghoub-Zadeh, Z., & Schuster, B. (2000). Understanding individual differences in word recognition skills of ESL children. *Annals of Dyslexia, 50* (1), 121-154.

(16) Geva, E., & Yaghoub Zadeh, Z. (2006). Reading efficiency in native English-speaking and English-as-a-second-language children: The role of oral proficiency and underlying cognitive-linguistic processes. *Scientific Studies of Reading, 10* (1), 31 - 57.

(17) Gough, P. B., & Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6-10.

(18) Grabe, W. (2009). *Reading in a second Language: Moving from Theory to Practice*. New York, NY: Cambridge University Press.

(19) Hocking, R. R. (1996). *Methods and applications of linear models: Regression and analysis of variance*. New York: John Wiley.

(20) Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2* (2), 127-160.

(21) Joshi, R. M. (1999). A diagnostic procedure based on reading component model. In J. Lundberg, F. E. Tonnessen & I. Austad (Eds.), *Dyslexia: Advances in theory and practice* (pp. 207-219). Dordrecht, Holland: Kluwer Academic Publishers.

(22) Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simple View of Reading made a little more complex. *Reading Psychology, 21*, 85-97.

(23) Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-36.

(24) Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16* (1), 33-51.

(25) Lesaux, N., Koda, K., Siegel, L., & Shanahan, T. (2006). Development of literacy. In D. August & T. Shanahan

(Eds.), *Developing literacy in second-language learners* (pp. 75-122). Mahwah, NJ: Erlbaum.

(26) Limbos, M. M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities, 34* (2), 136-151.

(27) Linacre, J. M. (2007). A user's guide to WINSTEPS: Rasch-model computer program. Chicago: MESA Press.

(28) Linacre, J. M. (n.d.-a). Winsteps manual (Dimensionality: contrasts & variances) Retrieved 07/15, 2010, from http://www.winsteps.com/winman/principalcomponents.htm

(29) Linacre, J. M. (n.d.-b). Winsteps manual (Dimensionality: Contrasts & variances) Retrieved 08/09, 2010, from http://www.winsteps.com/winman/principalcomponents.htm

(30) Linacre, J. M. (n.d.-c). Winsteps manual (Table 23.0 Variance components scree plot for items) Retrieved 08/06, 2010, from http://www.winsteps.com/winman/table23_0.htm

(31) Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal, 75* (2), 196-204.

(32) Maeng, U.-K. (2006). Comparison of L2 listening and reading comprehension strategies: A case study of three middle school students. *The Journal of Curriculum & Evaluation, 9* (2), 471-500.

(33) Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning, 11* (2), 323-348.

(34) Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111* (1), 172-175.

(35) Murphy, J. M. (1996). Integrating listening and reading instruction in EAP programs. *English for Specific Purposes, 15* (2), 105-120.

(36) Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31* (7), 9-13.

(37) Norušis, M. J. (2008). *SPSS16.0 statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.

(38) Osborne, J. W., & Waters, E. (n.d.). Four assumptions of multiple regression that researchers should always test Retrieved 09/30, 2010, from http://pareonline.net/getvn.asp?v=8&n=2

(39) Park, G.-P. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals, 37* (3), 448-458.

(40) Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

(41) Raîche, G. (n.d.). Critical eigenvalue sizes in standardized residual principal components analysis (PCA) Retrieved 08/09, 2010, from http://www.rasch.org/rmt/rmt191h.htm

(42) Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

(43) Royer, J. M. (1985). Reading from the perspective of a biological metaphor. *Comtemporary Educational Psychology, 10*, 150-200.

(44) Sanders, D. A. (1977). *Auditory perception of speech: An introduction to principles and problems*. Englewood Cliffs, NJ: Prentice-Hall.

(45) Sinatra, G. M. (1990). Convergence of listening and reading process. *Reading Research Quarterly, 25* (2), 115-130.

(46) Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2* (3), 281-311.

(47) SPSS 15.0 for Windows [Computer software] (2007). Chicago: SPSS, Inc.

(48) Statistics Solutions (n.d.). Scatterplot: An assumption of regression analysis Retrieved 09/30, 2010, from http://www.statisticssolutions.com/methods-chapter/statistical-tests/scatterplot/

(49) Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

(50) Sticht, T. G., Beck, L. J., Hank, R. N., Kleiman, G. M., & James, J. H. (1974). *Auding and reading: A developmental model*. Alexandria, VA: Human Resources Research Organization.

(51) Thompson, I. (1995). Assessment of second/foreign language listening comprehension. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 31-58). San Diego, CA: Dominie

Press.
(52) Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal, 90* (i), 6-18.
(53) Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire: Development and validation. *Language Learning, 56* (3), 431-462.
(54) Wade-Woolley, L., & Siegel, L. S. (1997). The spelling performance of ESL and native speakers of English as a function of reading skill. *Reading and Writing, 9* (5), 387-406.

(Associate Professor, The International Center, Kyoto University)

京大生の英語長文聴解力

青谷　正妥

要旨

京大生の長文聴解力（TOEFL iBT 形式）を、15 のテスト群で定量的に検証した。Rasch analysis で線形化した様々な英語力の指標を変数として用い、correlation matrix, multiple linear regression, principal component analysis で指標間の相関と聴解力への貢献度の明確化を試みた。音素や単語レベルでのインプット処理がより重要な役割を果たす短文聴解（TOEFL PBT/ITP 形式）に比して、長文聴解力は全体的な意味の理解が優先するので、むしろ読解力と共通部分が多く、実際、読解と Listening cloze だけで、長文聴解力全体の 70% 近くが説明できた。更に文法・構文の認識・判断力が、京大生の場合には長文・短文の聴解力を大きく左右しない事も明らかとなった。読解力が聴解力に繋がる図式は、日本人英語学習者にとっては、ある意味で朗報であろうが、テスト作成者は、長文聴解のみでは音を捉えるという意味での聴解力の「純粋な聴解」の部分は測れないという事実に、十分に留意する必要がある。

（京都大学国際交流センター・准教授）