Research Paper

Kyoto University Students' Grammatical Abilities and Listening Comprehension as Measured by the TOEFL PBT and GRE

Masayasu Aotani

Abstract

Kyoto University students' grammatical and syntactic skills, as well as their short listening comprehension skills, were compared with the corresponding world averages as measured by TOEFL. Grammatical and syntactic skills were operationalized as the structure (gap filling) and written expression (grammatical error detection) section scores from the TOEFL PBT (Paper-Based-Test) and the scores from the GRE's gap filling questions. Short listening comprehension skills were operationalized as the scores from the TOEFL PBT's short listening section. The average English learner, by our definition, has a probability of success on each item which is equal to the world average made available by Educational Testing Service (ETS). Taking advantage of the Rasch-type symmetry between person abilities and item difficulties, the test items were regarded as examinees taking two tests; made of Kyoto University students and all the examinees around the world. This novel symmetric view made extensive correlation studies possible. While Kyoto University students excelled at TOEFL's gap filling task, there was no statistically significant difference in the scores for grammatical error detection, contradicting the common belief that Japanese students have superior grammatical abilities. This may be due to insufficient procedural skills to use declarative grammar knowledge fluently. Lack of procedural abilities may partially explain the markedly poorer performance in the second half of the GRE's gap filling test as well.

【Keyword】 grammatical knowledge, listening comprehension, Japanese learners of English, Rasch analysis, TOEFL, GRE

1. Introduction

There is a common belief that Japanese students excel at tasks requiring grammatical and syntactic skills but are weak in listening comprehension (Ikeda, n.d.; Saegusa & Gay, 1988). I examined the validity of this claim as reflected in the TOEFL PBT (listening, structure, written expression) and GRE (gap filling) scores. Scores for Kyoto University students on the listening comprehension section and structure and written expression section of the TOEFL, and the GRE gap filling section were compared with the results published by Educational Testing Service (ETS) for all examinees (1995, 2003). This study was partly motivated by the performance of Japanese

examinees on the TOEIC reading and listening sections. Despite the common belief that Japanese people's English reading comprehension skills are better than their corresponding listening skills, many Japanese examinees get a higher score on the listening section of the TOEIC (Kaufmann, n.d.). It does not seem that one can make such a categorical statement about the grammatical and listening abilities of Japanese examinees. It may vary as a function of the type of instrument used. It may depend on the amount of time allowed. Or, it may simply be that such a belief is wrong. Common beliefs should be treated exactly as such; that is, they are beliefs and not proofs. This applies to our belief about Japanese people's grammatical/syntactic and listening skills as well. The aim here is to examine this belief and determine its validity, at least in the particular context of the current project. I used the TOEFL PBT because the TOEFL iBT (internet-Based-Test), the current version of TOEFL, does not have a grammar section, and the TOEFL iBT's section scores for Kyoto University students and the world averages are more or less evenly matched, as shown in Table 1. Furthermore, the lack of sufficient data due to the short history of the TOEFL iBT necessitated the exclusive use of the TOEFL PBT.

Table 1. ***Average TOEFL iBT Scores for 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. Due to rounding, the total of the section averages do not match the average total score. The All Japanese average was from 2007, while Kyoto University's average was a cumulative average over three years (2007-2010) and only scores submitted voluntarily are included. For the sake of comparison, the average TOEFL iBT scores for all admitted international graduate students at Princeton University are: Reading 29, Listening 28, Speaking 24, Writing 27, and Total 108.

This research was a subset of a larger project, where 112 Japanese-speaking Kyoto University students' listening comprehension was studied with a battery of 15 tests in order to find their proficiency profile, focusing on the subskills involved. In particular, systematic comparisons of various potential contributing factors/subskills were made between long/holistic listening and short listening in the parent study. See Aotani (2009, 2011a, 2011b) for more details of this parent project.

It is with listening that humans start their linguistic activities, either as a newborn or even before birth (H. D. 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), with the other three being speaking, reading, and writing. This is clearly illustrated by the much smaller number of publications on listening than on the other receptive skill of reading. One approach to characterizing listening comprehension is to decompose the listening skill into component subskills. One can study both the underlying psycholinguistic parameters such as working memory capacity, and the more traditionally recognized subskills of listening, which include word recognition and grammatical/syntactic knowledge. This paper examines the latter subskills of Kyoto University students. In particular, detailed quantitative studies were conducted

on the relative strengths of their short listening comprehension and grammatical/syntactic judgment skills.

2. Materials and methods

2.1 Participants

The initial participants of this study were 179 Kyoto University students, mostly freshmen, who were taking the author's English classes, which were offered in the spring semester of 2010. Their mother tongue was Japanese. The actual number of examinees differed from one test to another due mainly to absences, so only the 112 students who did not miss an examination in my statistical computations. These students all received at least six years of English education in junior and senior high school, amounting to about 3,000 hours of exposure to English, with an emphasis on reading and writing. They also passed Kyoto University's highly competitive entrance examination, whose English section consisted exclusively of translations from English to Japanese and vice versa.

2.2 Instruments

I used the first 30 questions from the TOEFL PBT's Listening Comprehension section (the Short Listening Test: SLT), 30 questions from the TOEFL PBT's Structure section (the Gap Filling Test: GFT), and 25 questions from the TOEFL PBT's Written Expression section (the Grammatical Error Detection Test: GED), for which item-wise average success rates for all examinees were made available by ETS. I also included 35 questions from the gap filling section of the Graduate Record Examinations (the GRE Gap Filling Test: GGT) for comparison. The characteristics of these tests are presented in Table 2. Examples of test items can be found in Appendices A through D.

Table 2. ***Tests and Their Characteristics***

Test Name (Abbreviation)	Short Description of Content and Purpose
Short Listening Test (SLT)	A short conversation with two turns followed by a multiple-choice question To measure sentence-level listening comprehension
Grammatical Error Detection Test (GED)	Finding the segment containing a grammatical error from four underlined segments in each sentence To measure receptive grammatical judgment
Gap Filling Test (GFT)	Selecting the correct expression to fill a gap in each sentence (This test is for nonnative speakers of English.) To measure any or all of the following: sentence- or subsentence-level semantic/syntactic awareness, sensitivity to coherence or consistent flow of ideas, and the main ideas of the passage (Alderson, 2000, pp. 207-211)
GRE Gap Filling Test (GGF)	Selecting the correct expression to fill a gap or two in each sentence (This test is designed for native speakers of English.) To measure sentence-level semantic/syntactic awareness and sensitivity to coherence (Alderson, 2000, pp. 207-211)

2.3 Methods

As the data provided by the ETS is the average success rate for each item, the same was computed for the 112 participants of this investigation, and Pearson correlation studies of the success rates between the different groups of examinees were conducted first. This was followed by analysis of the mean scores for different tests and different groups. Independent-samples *t*-test and paired-samples *t*-test were employed to examine the significance of the differences among the means. I used the raw scores for all of my analyses because item-by-item data for individual examinees are not made available by ETS, making scale adjustment and linearization procedures such as Rasch analysis impossible.

3. Results

3.1 Correlation studies

Table 3. ***Item-Wise Success Rate: ETS Data vs. the Study Participants***

	SLT		GED		GFT		GGF	
Item	ETS	Participants	ETS	Participants	ETS	Participants	ETS	Participants
1	.87	.79	.94	.95	.93	.92	.90	.52
2	.81	.60	.86	.70	.84	.96	.82	.33
3	.89	.75	.80	.77	.93	.99	.81	.21
4	.75	.59	.88	.96	.83	.99	.77	.46
5	.85	.62	.87	.86	.78	.88	.70	.14
6	.86	.75	.85	.78	.82	.94	.62	.20
7	.79	.49	.80	.93	.76	.86	.28	.30
8	.91	.72	.81	.82	.67	.90	.84	.10
9	.84	.62	.84	.74	.65	.91	.86	.52
10	.88	.48	.72	.80	.66	.81	.87	.42
11	.83	.77	.76	.94	.54	.78	.80	.53
12	.71	.50	.72	.88	.62	.85	.74	.48
13	.86	.81	.71	.89	.41	.50	.71	.44
14	.79	.32	.70	.65	.45	.54	.68	.07
15	.81	.50	.62	.84	.95	.99	.79	.39
16	.84	.52	.67	.75	.89	.98	.95	.39
17	.57	.40	.65	.57	.80	.96	.88	.41
18	.57	.28	.57	.68	.86	.99	.75	.81
19	.71	.40	.59	.80	.89	.97	.56	.42
20	.53	.34	.57	.61	.80	.92	.57	.11
21	.61	.27	.41	.46	.81	.94	.42	.10
22	.54	.32	.52	.43	.88	.91	.76	.44
23	.55	.32	.57	.46	.71	.89	.70	.45
24	.53	.42	.37	.38	.68	.66	.57	.30
25	.62	.52	.36	.51	.63	.90	.72	.30
26	.53	.29			.43	.70	.63	.26
27	.55	.34			.34	.33	.55	.07
28	.72	.38			.36	.55	.52	.13
29	.30	.11			.32	.45	.93	.45
30	.41	.28			.86	.94	.91	.48
31							.79	.15
32							.69	.12
33							.68	.14
34							.54	.20
35							.58	.08
Total	21.03	14.48	17.16	18.14	21.10	24.91	24.89	10.92

Note. SLT = Short Listening Test; GED = Grammatical Error Detection; GFT = Gap Filling Test; GGF = GRE Gap Filling Test. The Short Listening Test and the Gap Filling Test had 30 items, the GED 25 items, and the GGF 35 items. The second row from the bottom shows the mean values for the total scores.

Item-wise success rates for each group of examinees are shown in Table 3. As any English teacher in Japan would expect, Kyoto University students' average scores on the Short Listening Test (SLT) and the GRE Gap Filling Test (GGF) appear clearly lower than the average for all the examinees. On the other hand, the differences are small for the Grammatical Error Detection Test (GED) and the Gap Filling Test (GFT). I will examine the statistical significances of all these differences in this section.

Correlations between the scores of the participants and the data provided by ETS are shown in Table 4. The first three (SLT, GED, and GFT) are between Kyoto University students and the other examinees from around the world who are also nonnative speakers of English. For the GRE (GGF), however, most examinees are either native speakers or advanced learners of English. This seems to be well reflected in the correlation coefficients.

Table 4. ***Score Correlations between the ETS Data and the Scores of the Participants***

Test	SLT	GED	GFT	GGF
Correlation	.837	.792	.908	.481

Note. SLT = Short Listening Test; GED = Grammatical Error Detection Test; GFT = Gap Filling Test; GGF = GRE Gap Filling Test.
$p < .01$ for all correlations

3.2 Comparison of Means

Both a paired-samples *t*-test and a two independent-samples *t*-test were conducted to examine whether the mean scores for the study participants and those from the ETS-provided data were significantly different. Because individual scores are unavailable for the ETS data, it is best to regard the same set of items as being answered twice by different populations. In this sense, paired-samples *t*-tests are appropriate. However, an independent-samples *t*-test is more stringent in deciding the significance of the difference between two means, and the paired-samples design is not as effective when the correlation is close to .50 (Norušis, 2008, p. 139) as is the case with the GRE Gap Filling Test. For all the two-independent samples *t*-tests conducted for mean comparison in this chapter, the combined sample size of the two groups was greater than 40. Therefore, the normality assumption for the two populations is not a concern for the two independent-samples *t*-tests (Norušis, 2008, pp. 140-141). Note here that the sample size in this context is the number of test items and not the number of test takers. For the paired-samples *t*-test, the requirements are for the differences and not the original observations. This is because a paired-samples *t*-test is nothing more than a one-sample *t*-test on the differences (Norušis, 2008, p. 138). With the sample sizes falling in the range between 15 and 40, the set of differences between two measurements should not have any outliers or significant skewness (Norušis, 2008, p. 135). Hence, it is necessary to check the descriptive statistics for the dataset containing the differences between the two measurements. Relevant descriptive statistics for the differences between the ETS and Kyoto University data are presented in Table 5.

Data points outside the range (Mean – 3SD, Mean + 3SD) were regarded as outliers. One outlier was identified for the GED + GFT variable. However, the score was only 1 point above the acceptable limit, so the datum was retained. According to Brown (1997, p. 20), "Values of 2

standard errors of skewness (*ses*) or more (regardless of sign) are probably skewed to a significant degree." Applying this criterion, no data set was skewed significantly.

Table 5. ***Descriptive Statistics for the Differences between ETS and Kyoto University Participants***

Variable	*N*	*M*	*SD*	Skewness	SES	Kurtosis	SEK
Short Listening Test	30	.218	.101	.444	.427	-.011	.833
Grammatical Error Detection	25	-.040	.110	.003	.464	-1.071	.902
Gap Filling Test	30	-.127	.082	-.117	.427	-.538	.833
GRE Gap Filling Test	35	.399	.170	-.714	.398	1.167	.778
GED+GFT	55	-.088	.104	.297	.322	-.446	.634
SLT+ GED+GFT	85	.020	.179	.382	.261	-.748	.517

Note. These are descriptive statistics for the set of differences. For example, for Item 1 of SLT, the average success rate for the ETS data is .87 and that for the study participants is .79 as shown in Table 2. Hence, the value for the first item is .87 – .79 = .08. There are 30 differences for the Short Listening Test, each corresponding to one of the 30 items on the test, calculated in this manner.

The results of the paired-samples *t*-test and the two-independent-samples *t*-tests are presented in Table 6. As expected, the differences for the Short Listening Test (SLT) scores and the GRE Gap Filling Test (GGF) scores were clearly significant at $p < .0005$ for both paired and independent tests. In addition, the difference for the Gap Filling Test (GFT) was also significant. However, it turned out Kyoto University participants' grammatical error detection skills (GED) were not significantly different from that of all the examinees despite the common belief about Japanese college students' superior grammatical ability.

Table 6. ***Comparison of Means: Independent-Samples and Paired-Samples Tests***

Tests	SLT		GED		GFT		GGF	
Means	ETS	Kyoto	ETS	Kyoto	ETS	Kyoto	ETS	Kyoto
	21.03	14.48	17.16	18.14	21.10	24.91	24.89	10.92
Paired	.000 (***)		.084 (*ns*)		.000 (***)		.000 (***)	
Indep.	.000 (***)		.417 (*ns*)		.011 (*)		.000 (***)	

Note. SLT = Short Listening Test; GED = Grammatical Error Detection Test; GFT = Gap Filling Test; GGF = GRE Gap Filling Test.
* $p < .05$ *** $p < .0005$ *ns* means not significant at $p < .05$

More detailed results are presented below in Tables 7 through 11. For independent-samples *t*-tests, only the results for the SLT and GED are shown.

Table 7. ***Group Statistics (SLT): Independent-Samples Test***

		N	*M*	*SD*	*SEM*
SLT	ETS	30	.7010	.16236	.02964
	Kyoto	30	.4827	.18587	.03394

Table 8. ***Independent Samples t-test (SLT)***

		Levene's Test		t-test for Equality of Means						
									95% Confidence Interval	
	Equal variances	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SLT	assumed	.330	.568	4.845	58	.000	.21833	.04506	.12813	.30852
	not assumed			4.845	56.970	.000	.21833	.04506	.12810	.30856

Table 9. ***Group Statistics (GED): Independent-Samples Test***

		N	*M*	*SD*	*SEM*
GED	ETS	25	.6864	.16230	.03246
	Kyoto	25	.7257	.17646	.03529

Table 10. ***Independent Samples Test (GED)***

		Levene's Test		t-test for Equality of Means						
									95% Confidence Interval	
	Equal variances	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
GED	assumed	.262	.611	-.819	48	.417	-.03928	.04795	-.16789	.08933
	not assumed			-.819	47.668	.417	-.03928	.04795	-.16793	.08937

Table 11. ***Paired Samples Test***

	Paired Differences					t	df	Sig. (2-tailed)
				95% Confidence Interva				
	Mean	SD	SEM	Upper	Lower			
SLT: ETS - Kyoto	.21767	.10136	.01851	.17982	.25551	11.762	29	.000
GED: ETS - Kyoto	-.04125	.11183	.02283	-.08847	.00597	-1.807	24	.081
GFT: ETS - Kyoto	-.12700	.08154	.01489	-.15745	-.09655	-8.531	29	.000
GGF: ETS - Kyoto	.39914	.16964	.02867	.34087	.45742	13.920	34	.000

Because the Grammatical Error Detection Test and the Gap Filling Test together are hypothesized to measure grammatical knowledge and syntactic awareness, and the Short Listening Test (SLT), the Grammatical Error Detection Test (GED), and the Gap Filling Test (GFT) form a core part of the now obsolete paper-based TOEFL (TOEFL PBT), we also computed correlation coefficients for the combined GED and GFT scores and for the total of all three scores. These are presented in Table 12, along with the mean total scores for the participants in this study and those from the ETS data. When the mean scores for the participants and ETS data were compared, the difference was significant at $p < .05$ for GED + GFT and not significant at $p < .05$ for Short Listening Test + GED + GFT as indicated in the last two rows. Both the independent-samples and the paired-samples *t*-tests were conducted.

Table 12. ***Means for and the Correlations between Combined Scores***

Test		GED + GFT		SLT + GED + GFT	
Correlation		.837		.650	
Mean Total Score		ETS	Kyoto	ETS	Kyoto
		38.26	43.05	59.29	57.53
Sig.	Paired	.000 (***)		.300 (*ns*)	
	Independent	.013 (*)		.513 (*ns*)	

Note. *** $p < .0005$ * $p < .05$ *ns* means not significant. The last row labeled "Sig." means the significance of the differences between the two mean scores; ETS and Kyoto.

3.3 Two Halves of the GRE Gap Filling Test

A visual inspection revealed that Kyoto University students' performance on the GRE Gap Filling Test was noticeably worse on the second half of the test. Furthermore, the correlation coefficient between the total scores on the first 18 questions and the last 17 questions (Questions 19 through 35) was only .272. A repeated-samples *t*-test showed that the means for the first and second halves were significantly different at $p < .0005$. As the GRE Gap Filling Test was designed to measure processing speed as well, the average time allowed for each item was deliberately set at 50 seconds; this is the same amount of time as is allowed for the actual GRE and is shorter than what is comfortable for most participants. Though almost all the students managed to complete the test, it is possible that they were running out of time and rushed towards the end of the test. In order to investigate this possibility, the test was divided into the first 18 questions (Questions 1 to 18) and the second 17 questions (Questions 19 to 35), and the mean success rates per item for the two halves were compared. This was done for both Kyoto University students and for the data provided by ETS. The results are shown in Table 13.

Table 13. ***Comparison of Performances on the First and the Second Halves of the GGF***

	Questions 1 to 18		Questions 19 to 35	
Mean Success Rate	Kyoto	ETS	Kyoto	ETS
	.3735	.7650	.2471	.6541

Note. The difference between the mean success rates for the first and the second halves was significant at $p < .05$ for both Kyoto University students and the ETS data. The difference between Kyoto University students and the ETS data was significant at $p < .01$ both for the first half (Questions 1 to 18) and the second half (Questions 19 to 35).

The significance levels of the pair-wise differences among the four means were computed. For simplicity, I denote the mean of:

the first 18 questions for Kyoto University students by M1K,

the last 17 questions (Questions 19 through 35) for Kyoto University students by M2K,

the first 18 questions from the ETS data by M1E,

and

the last 17 questions from the ETS data by M2E.

Table 14 summarizes the significance of the pair-wise differences among the means.

Table 14. ***Significances of Pair-wise Differences among the Means of Two Halves of the GRE Gap Filling Test***

	M2K	M1E	M2E
M1K	.032[a]	.0005[b] .0005[a]	-----
M2K	-----	-----	.0005[b] .0005[a]
M1E	-----	-----	.028[a]

[a]Independent-samples *t*-test
[b]Paired-samples *t*-test

Performance was significantly worse at $p < .05$ on the second half of the GRE Gap Filling Test for both groups of students. In order to place item difficulties and person abilities on the same linearized scale, the Rasch method was employed. My analysis of the GRE Gap Filling Test for the Kyoto University students revealed the following differential between the item difficulties of the first 18 and the remaining 17 items. Recall that item by item performance data for individual examinees are not published by ETS, and Rasch analysis could not be performed for them, limiting the scope of this part of the investigation to Kyoto University participants.

It appears via visual inspection of the item-person map (Figure 1) that there is a larger cluster of more difficult items in the second half of the test that is consistent with the observed lower mean success rate for that half. Numerical computation confirmed that the mean item difficulty measure for Questions 1 through 18 was -.35, while that for Questions 17 through 35 was .37. Because these means are in logits, the ratio of raw difficulties defined as

$$\frac{1-P_{0i}}{P_{0i}},$$

where 0 stands for the person selected as the reference and P_{0i} is the probability that the reference person will answer item i correctly (E. V. Smith, Jr. & Smith, 2004, p. 11; 2007; R. M. Smith, 1992, 2003), can be obtained as follows. We will denote the raw difficulty corresponding to -.35 and .37 by $d-$ and $d+$ respectively.

$$\ln(d+)-\ln(d-)=0.37-(-0.35)=0.72 \Rightarrow \ln\frac{d+}{d-}=0.72$$

$$\Rightarrow \frac{d+}{d-}=e^{0.72}\approx 2.05$$

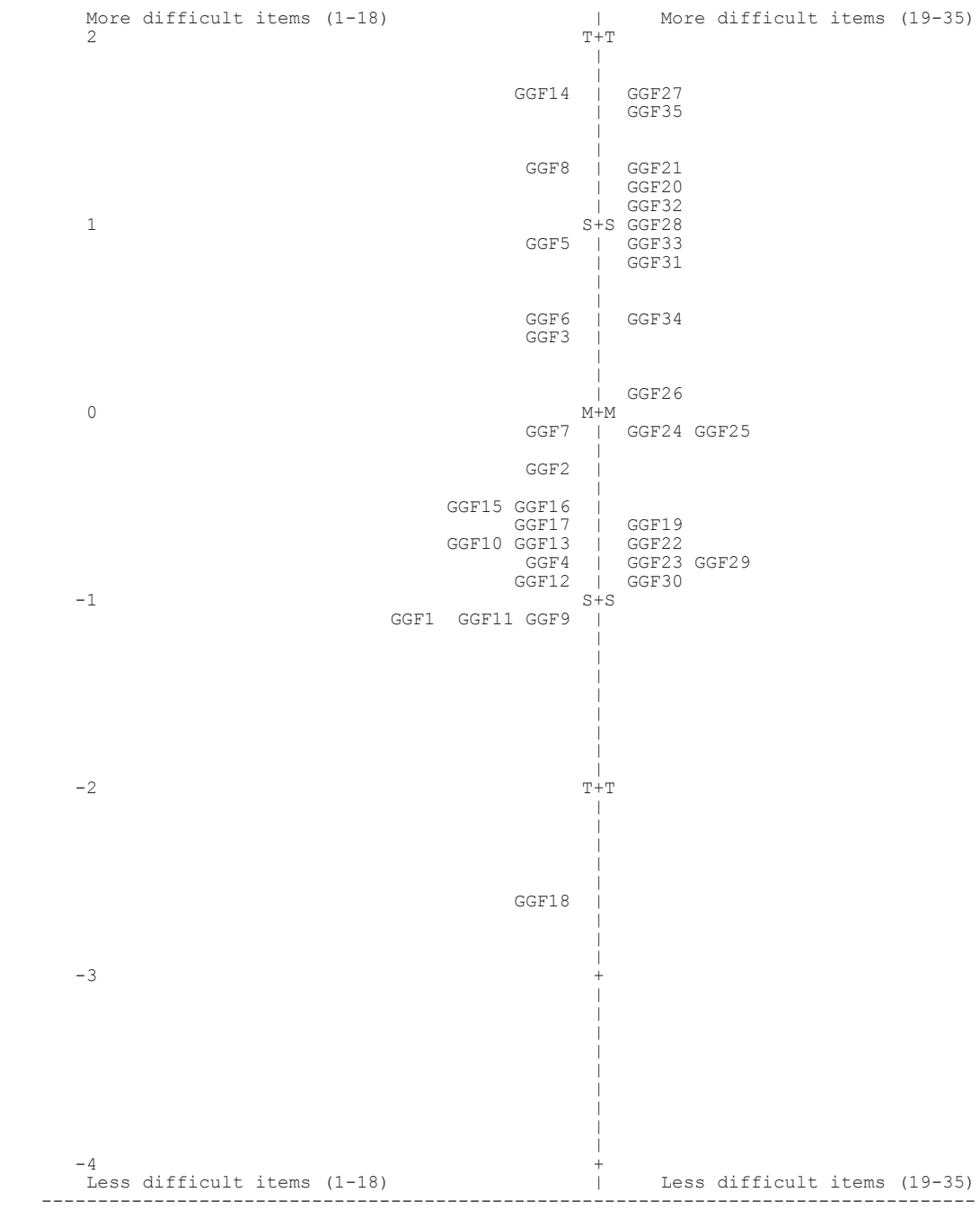

Figure 1. Item Difficulties for the First and Second Halves of the GRE Gap Filling Test.

Hence, in terms of the raw difficulty, Items 19 through 35 are about twice as difficult as Items 1 through 18 on average. More specifically, it is possible to compute the probability x that the reference person 0 will answer an item of difficulty measure .37 correctly and the probability y that person 0 will answer an item of difficulty measure -.35 correctly. For x, we have

$$\ln\frac{1-x}{x}=0.37\Rightarrow\frac{1-x}{x}=e^{0.37}\Rightarrow x=\frac{1}{1+e^{0.37}}=0.409\,.$$

Similarly, we get

$$y=\frac{1}{1+e^{-0.35}}=0.587\,.$$

Note that $x < y$ as the probability for correctly answering the more difficult item, of measure .37 in this case, should be lower.

From these considerations, it is clear that there is a significant difference between the first and the second halves in terms of the Rasch difficulty measure. However, it is not clear whether this difference is due exclusively to genuine item difficulty, that is, the intrinsic difficulty of the item, or such external factors as time shortage and fatigue in the second half. There was a 34% decline in the success rate on the second half for the Kyoto University students as opposed to a 14% decrease as measured for the ETS data. Whether this difference derives from different processing speeds or different proficiency levels is not clear, either. It could be that there is a threshold difficulty level for the participants in this study, who are all nonnative speakers, beyond which the problems suddenly become far more difficult as though there is a linguistic phase transition at that point, causing the observed performance shift.

4. Discussion

From Tables 3, 4, 6, 12, 13, and 14, the following general characteristics emerged.

1. As expected, the students' performance profile is much more like that of other learners of English than native speakers and advanced learners of English. This is clear when one compares their performances on the tests for nonnative speakers (SLT, GED, GFT) and on the GRE Gap Filling Test, where the questions were written with native speakers in mind, with the data provided by the ETS. The correlations shown in Table 3 clearly indicate that the parts of Kyoto University students' proficiency profile measured by SLT, GED, GFT, and the GGF are much closer to that for other nonnative speakers than to native speakers and advanced learners. The correlation ranged from .792 to .908 when compared with other nonnative speakers, but was only .481 when compared with native speakers and advanced learners of English. A comparison of the means fully corroborated this finding and interpretation.

2. The participants' scores on the Short Listening Test clearly lag behind the average of all test takers. A difference of 6.55 points out of a total score of 30 is significant, as shown by the *t*-test ($p < .0005$) . Parenthetically, it has been well known to the teachers at Kyoto University that the students excel at more holistic listening such as the listening comprehension section of TOEFL iBT, where a good understanding of the main ideas of longer texts is required.

3. Despite a popular belief about Kyoto University students' high grammatical competence, their receptive grammatical judgment, as measured by the error spotting tasks on the GED, is only as good as that for an average test taker, according to the ETS data. There was no significant difference between the mean scores of the Kyoto University students and all test takers ($p = .417$).

4. Kyoto University students excelled at TOEFL PBT's gap filling task. The difference between the mean scores of the study participants and the ETS data was significant at $p < .0005$ for the repeated-samples *t*-test and at $p < .011$ for the independent-samples *t*-test.

5. As a result of findings 2, 3, and 4 above, with the clear advantage of the Kyoto University students in the Gap Filling Test and the even clearer advantage of the average TOEFL PBT examinee on the Short Listening Test apparently cancelling each other out, the mean scores for the Kyoto University students and the average test taker are not significantly different for the combined score of the Short Listening Test + Grammatical Error Detection Test + Gap Filling Test ($p = .300$ for the repeated-samples *t*-test and .513 for the independent-samples *t*-test) as shown in Table 11. Though the difference was significant even based on the independent-samples *t*-test with $p = .013$ for GED + GFT, this was to be expected as the average scores on the GED (Grammatical Error Detection Test) and GFT (Gap Filling Test) were both higher for the students in this study. It is well-known that Japanese test takers do well on the Grammar and Structure section of the TOEFL PBT, which consists of GED and GFT type questions, and the students in this study were no exceptions. Before leaving this issue, I will make a note of the fact that this kind of cancellation, between the sections where the study participants excelled and the sections where the average examinee worldwide excelled, is also observed for the TOEFL iBT. Despite the weakness of Kyoto University students in the Speaking section, their average total score of 78 is the same as the average for all the examinees worldwide. This is due mainly to their high scores on the Reading section (see Table 1).

6. Kyoto University students' performance on the second half of the GRE Gap Filling Test was significantly worse ($p = .032$) than on the first half. Though native speakers and advanced learners of English also did significantly worse ($p = .028$) on the second half of the test, their success rate did not decline nearly as much. Further investigations are necessary to find the cause of this difference. Processing speed, difficulty threshold, and fatigue level are all potential sources of this variation. I strongly suspect processing speed is the main cause.

5. Conclusion

There is a clear indication that the students attending Kyoto University, and most likely other college students in Japan, should work on their short listening skills, such as sound perception and word recognition. Short listening requires more efficient bottom-up processing than longer listening, which provides more context for the listeners to rely on. This weakness is the primary reason why their performances on tests like the TOEFL PBT are mediocre. With regard to grammatical skills, it appears that the students fail to apply their declarative knowledge to correctly answer the questions. Therefore, their numerically measured skill levels vary rather widely depending on which aspect/type of grammatical/syntactic knowledge is tapped by the instruments used, as well as how much processing speed is required.

References

(1) Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.
(2) Aotani, M. (2009). *Proficiency profile of Kyoto University students*. Unpublished study. The International Center, Kyoto University. Kyoto, Japan.
(3) Aotani, M. (2011a). *Factors affecting the holistic listening of Japanese learners of English.* Ed.D., Temple University, Philadelphia.
(4) Aotani, M. (2011b). Factors contributing to holistic listening of Kyoto University students: A preliminary study. *The International Center Research Bulletin, Kyoto University, 1*, 21-43.
(5) Brown, H. D. (1987). *Principles of language learning and teaching*. Englewood Cliffs, NJ: Prentice Hall.
(6) Brown, J. D. (1997). Skewness and Kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(1), 20-23.
(7) 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.
(8) Educational Testing Service. (1995). *GRE: Practicing to take the general test: Big book*. Princeton: Educational Testing Service.
(9) Educational Testing Service. (2003). *TOEFL: Test preparation kit: Workbook* (2nd ed.). Princeton: Educational Testing Service.
(10) Educational Testing Service. (2009). *The official guide to the TOEFL test* (3rd ed.). Princeton: Educational Testing Service.
(11) Ikeda, M. (n.d.). Teaching English to Japanese students Retrieved 12/25, 2010, from http://humanities.byu.edu/elc/Teacher/japanesestudents.html
(12) Jalongo, M. R. (2010). Listening in early childhood: An interdisciplinary review of the literature. *International Journal of Listening, 24*(1), 1-18.
(13) Kaufmann, S. (n.d.). TOEIC - One reason why Japanese struggle with English - they don't read in English. Retrieved 12/25, 2010, from http://humanities.byu.edu/elc/Teacher/japanesestudents.html
(14) Norušis, M. J. (2008). *SPSS16.0 statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.
(15) Robinshaw, H. (2007). Acquisition of hearing, listening and speech skills by and during key stage 1. *Early Child Development and Care, 177*(6 & 7), 661-678.

(16) Saegusa, Y., & Gay, C. W. (1988). Japanese students' English proficiency. *Waseda Journal of Human Sciences, 1*(1), 3-13.
(17) Smith, E. V., Jr., & Smith, R. M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.
(18) Smith, E. V., Jr., & Smith, R. M. (2007). *Rasch measurement: Advanced and specialized applications*. Maple Grove, MN: JAM Press.
(19) Smith, R. M. (1992). *Applications of Rasch measurement*. Maple Grove, MN: JAM Press.
(20) Smith, R. M. (2003). *Rasch measurement models: Interpreting WINSTEPS and FACETS output*. Maple Grove, MN: JAM Press.

Acknowledgment

I would like to sincerely thank Drs. David Beglar and Michael Linacre for their expert advice and guidance. It was only with their dedicated support that I was able to apply my knowledge of pure mathematics to the linguistic analyses of this paper.

(Associate Professor, The International Center, Kyoto University)

APPENDIX A
SHORT LISTENING TEST

短い会話を聞き、質問の答えを四つの選択肢から選んで下さい。

Listen to a short conversation and choose the right answer to the question from the four printed choices.

1. (woman) Thanks a lot! This scarf will be perfect with my blue jacket.
   (man) Made a good choice, did I?
   (narrator) What does the man mean?
   (A) He wants to know which scarf the woman chose.
   (B) He wants to know what color the jacket is.
   (C) He thinks he selected a nice scarf.
   (D) He thinks any color would go well with the jacket.

2. (woman) My cousin Bob is getting married in California and I can't decide whether to go.
   (man) It's a long trip, but I think you'll have a good time.
   (narrator) What does the man imply?
   (A) Bob has been married for a long time.
   (B) The woman should go to California.
   (C) He plans to go to the wedding.
   (D) He hasn't been to California for a long time.

3. (woman) Excuse me, could you bring me a glass of water, please?
   (man) Sorry, but I'm not a waiter.
   (narrator) What does the man mean?
   (A) He wants a glass of water.
   (B) He won't do as the woman asks.
   (C) He can't wait any longer.
   (D) He's looking for the waiter.

4. (man) Got the time?
   (woman) It's a little after ten.
   (narrator) What does the woman mean?
   (A) It's just past ten o'clock.
   (B) There's no time to talk.
   (C) She needs a little more time.
   (D) She has more than ten cents.

5. (man) You did an excellent job on that presentation.
   (woman) Thanks, I put a lot of time into it.
   (narrator) What does the woman mean?
   (A) She appreciates the man's help.
   (B) Her presentation was somewhat long.
   (C) She needed more time to prepare.
   (D) She worked hard on her presentation.

APPENDIX B
GRAMMATICAL ERROR DETECTION TEST

文法的な誤りを含む部分を、下線箇所から選んで下さい。各文に一つずつ誤りが有ります。
There is one grammatical mistake in each of the following sentences. Pick the underlined word/segment that contains a grammatical mistake.

1. Margaret Mead (A) studied many (B) different cultures, and she was one (C) of the first anthropologists to photograph (D) hers subjects.
2. Talc, (A) a soft mineral with a (B) variety of uses, (C) sold is in slabs or in powdered (D) form.
3. (A) During the 1870's iron workers in Alabama proved they (B) could produce iron by (C) burning iron ore with coke, (D) instead than with charcoal.
4. (A) Geologists at the Hawaiian Volcano Observatory (B) rely on (C) a number of instruments (D) to studying the volcanoes in Hawaii.
5. Underlying aerodynamics and (A) all other (B) branches of theoretical mechanics (C) are the laws of motion (D) who were developed in the seventeenth century.
6. (A) Was opened in 1918, the Phillips Collection (B) in Washington, D.C., was the first museum in the United States (C) devoted to modern (D) art.
7. A mortgage (A) enables a person (B) to buy property (C) without paying for it outright; thus more people are able to enjoy (D) to own a house.
8. (A) Alike ethnographers, ethnohistorians (B) make systematic observations, but they (C) also gather data from documentary and oral (D) sources.
9. Basal body temperature (A) refers to the (B) most lowest temperature of a (C) healthy individual (D) during waking hours.
10. (A) Research in the United States on acupuncture (B) has focused on (C) it use in (D) pain relief and anesthesia.
11. The Moon's (A) gravitational field (B) cannot keep atmospheric gases (C) from escape into (D) space.
12. (A) Although the pecan tree is chiefly (B) value for its fruit, its wood (C) is used extensively (D) for flooring, furniture, boxes, and crates.
13. (A) Born in Texas in 1890, Katherine Anne Porter produced three (B) collection of short (C) stories before (D) publishing her well-known novel Ship of Fools in 1962.
14. Insulation from cold, (A) protect against dust and (B) sand, and camouflage (C) are among the (D) functions of hair for animals.
15. (A) The notion that students are not sufficiently (B) involved in their education is one reason for the (C) recently surge of (D) support for undergraduate research.

APPENDIX C
GAP FILLING TEST

抜けている単語・表現を選んで下さい。

Fill in the blanks.

1. Dairy farming is ..... leading agricultural activity in the United States.
   A. a
   B. at
   C. then
   D. none

2. Although thunder and lightning are produced at the same time p light waves travel faster ....., so we see the lightning before we hear the thunder.
   A. than sound waves do
   B. than sound waves are
   C. do sound waves
   D. sound waves

3. Beef cattle ..... of all livestock for economic growth in certain geographic regions.
   A. the most are important
   B. are the most important
   C. the most important are
   D. that are the most important

4. The discovery of the halftone process in photography in 1881 made it ..... photographs in books and newspapers.
   A. the possible reproduction
   B. possible to reproduce
   C. the possibility of reproducing
   D. possibly reproduced

5. Flag Day is a legal holiday only in the state of Pennsylvania, ... Betsy Ross sewed the first American flag.
   A. which
   B. where
   C. that
   D. has

6. ..... vastness of the Grand Canyon, it is difficult to capture it in a single photograph.
   A. While the
   B. The
   C. For the
   D. Because of the

APPENDIX D
GRE GAP FILLING TEST

以下の各文には一つか二つ空白があり、単語や単語群が抜けています。選択肢 A~E より、最もこの文にふさわしいものを選んでください。

Each sentence below has one or two blanks, each blank indicating that something has been omitted. Beneath the sentence are five lettered words or sets of words. Choose the word or set of words for each blank that *best* fits the meaning of the sentence as a whole.

1. Nonviolent demonstrations often create such tensions that a community that has constantly refused to ..... its injustices is forced to correct them: the injustices can no longer be ..... .
   (A) acknowledge. .ignored
   (B) decrease. .verified
   (C) tolerate. .accepted
   (D) address. .eliminated
   (E) explain. .discussed

2. Since 1813 reaction to Jane Austen's novels has oscillated between ..... and condescension; but in general later writers have esteemed her works more highly than did most of her literary ..... .
   (A) dismissal. .admirers
   (B) adoration. .contemporaries
   (C) disapproval. .readers
   (D) indifference. .followers
   (E) approbation. .precursors

3. There are, as yet, no vegetation types or ecosystems whose study has been ..... to the extent that they no longer ..... ecologists.
   (A) perfected. .hinder
   (B) exhausted. .interest
   (C) prolonged. .require
   (D) prevented. .challenge
   (E) delayed. .benefit

4. Under ethical guidelines recently adopted by the National Institutes of Health, human genes are to be manipulated only to correct diseases for which ..... treatments are unsatisfactory.
   (A) similar
   (B) most
   (C) dangerous
   (D) uncommon
   (E) alternative

TOEFL PBT と GRE のスコアから見た京大生の文法理解と聴解力

青谷　正妥

要旨

京大生の文法・構文の知識を TOFL PBT の文法問題と穴埋め問題、更に GRE の穴埋め問題で測定し、同時に短文の聴解力を TOEFL PBT の聴解セクションの最初の 30 問で測定した。結果を Educational Testing Service が発表している全受験者のスコアと比較検討したところ、穴埋め問題では京大生が上回った。しかし、文法問題では、文法に強いはずの日本人学生の成績と世界平均との間に統計的に有意な差異は認められなかった。文法をふくめ、宣言的知識はあっても、それを利用するだけの技能はないという典型的な日本人の英語力の特性の現れであろうと思われる。手続き的知識が身につくに連れ、処理能力も高まるものであるが、未だ十分な手続き的知識の発達していない日本人学生には、GRE の穴埋め問題は与えられた時間が十分でなく、これが後半の出来が目立って悪かった理由ではないかと思われる。

（京都大学国際交流推進機構国際交流センター・准教授）