茨城大学教育学部紀要（教育科学）61 号（2012）207 － 215

Development of Listening Prochievement Tests for Third-Year Japanese Junior High School Students Studying English as a Foreign Language (Part II)

Hidetoshi SAITO*, Itsumi KINO**,and Takashi SAITO***

(Received November 25, 2011)

Abstract

The present study reports a process of equating of the two forms of a listening prochievement test for Japanese third-year junior high school students. The data came from two junior high schools. The main purpose of the report is to provide the details of the equating process of the two forms of the test.

Introduction: Purpose

The paper describes a process of equating of the two forms of a “prochievement” (proficiency and achievement) listening test, designed for measuring the English listening ability of the third-year students of junior high schools in Japan. Although the details of the setup and purposes are provided in Part I of this paper (Saito, Saito, & Kino, 2011), the information is briefly reiterated here for the readers of this paper.

The origin of the study had the purpose of measuring the learning gain of listening ability of the third-year junior high school students. Equating two test forms constitutes an important step towards measuring learning gains. The present research describes the equating procedure by using the data from the third to fifth administrations of the test (see Table 1).

Method

Participants

All test-takers were third-year junior high school students studying English as a foreign language in Japan. None of the students had extensive experience of staying overseas or being in intensive contact with native speakers of English. Version 2 was administered to two cohort groups

*Dpt. of English, College of Education, Ibaraki University, Mito 310-8512 Japan.
**Niihari Junior High School, Tsuchiura, 300-4115, Japan.
*** Ibaraki University Junior High School, Mito, 310-0056 Japan.

Table 1. Summary Information about Versions 2 and 3 of the Listening Test

Test Version	Version 2		Version 3	
Administration	3		4	5
Administration Time	Sept. 2008		Sept. 2010,	Feb. 2011
Number of Items	55		43	43
Number of Test-Takers	79 (A) 77 (B)		194 (A)	186 (B)
Groups	Group 2 (A)	Group 3 (B)	Group 1	Group 1

Notes. A and B indicates Forms A and B.

in a junior high school affiliated to a university, arbitrarily called Group 2 and Group 3 (Table 1). The school has a reputation of strong academic rigor. Version 3 was administered to students in a regular municipal junior high school (called Group 1). Based on the results of prefectural proficiency tests, called the Gakuryoku Shindan (academic diagnostic) Test, anecdotal evidence as well as the results of Part I of this paper (Saito, Saito, & Kino, 2011), Groups 2 and 3 excelled at listening ability compared to Group 1.

Instruments

Table 2 shows the overview of the data setup. The test (Version 3) consists of four sections: Section 1: 13 items of achievement questions based on the supplemental materials for the two nationally approved textbooks: *New Horizon English Course 3: Teacher's Manual* (Tokyo Shoseki, 2006) and *Total English 3 (Oyo-hatten ban waakushiito syuu)* (2006).

Section 2: 10 items of proficiency questions taken from the STEP test in which all response options did not appear on the test paper.

Section 3: 10 items of proficiency questions taken from the STEP test in which all response options were present on the test paper.

Section 4: 10 items of proficiency questions taken from the STEP test in which all response options were present on the test paper.

All questions of the STEP test papers are based on a Pre-2nd Grade STEP test preparation audio book *(Eiken Test in Practical English Proficiency Pre-2nd Grade,* 2008). The STEP test is an English proficiency test, which is widely recognized and used for high-stakes decisions, such as employment, promotion, and college admissions in Japan.

As seen in Table 1, the present study uses the data from three administrations of test Versions 2 and 3. The data of the second administration came from Groups 2 and 3, and items in these data were a subset of Version 3. That is, some items of Version 2 were revised and replaced with new

items (see Saito, Saito, & Kino, 2010), and thus only the items that remained in Version 3 were used for the present study. As seen in Table 2, three groups took different versions and forms of the test, but there were common items—anchor items—that overlapped. The concurrent calibration of the Rasch analysis makes it possible to link versions of the test using anchor items (see below).

Table 2. The Setup of the Test Data

		Achievement		Proficiency		
		Section 1		Section 2	Section 3	Section 4
Analysis		Anchor	Non-Anchor	Anchor	Non-Anchor	Non-Anchor
Form A	Group 1	4	9	10 (Form A)	10	10
	Group 2	3	7	10 (Form B)	10	10
Form B	Group 1	4	9	10 (Form B)	10	10
	Group 3	3	4	10 (Form A)	10	10

Notes. The number indicates the number of items in each Form. Group 1 took Version 3, while Groups 2 and 3 took Version 2 of the test.

There were two issues in the data. First, there were two independent analyses, (the first column of Table 2). Ideally, Groups in each analysis take the same test that contains identical anchor items. But as can be seen in the fifth column of Table 2, anchor items for Proficiency were nested within each group. That is, in the analysis of Form A, Group 2 did Section 2 from Form B, while Group 1 did Section 2 from Form A. Similarly, Group 3 did Section 2 from Form A, while Group 1 did Section 2 from Form B. Although the Rasch analysis—an item-response-theory analysis—has no problem in handling these types of gaps in the data for fair estimation of ability and item difficulty measures, it is true that there may be some bias because of this nestedness in the data (see Part I of this paper for more details about the Rasch analysis). However, this method allows us to double the number of anchor items from 10 to 20. In both Analyses, Section 2 of each Form was used for anchor items in addition to four items in Section 1. The increase in anchor items was a certainly advantage for the equating of the two Forms.

Another issue is that Group 1 repeatedly saw four achievement anchor items, and they were in both analyses. In other words, Group 1 took both Forms. It would be ideal for two independent groups to take different forms with overlapping anchor items. This was the case for Groups 2 and 3, where only one of the two forms was experienced. In the present study, the data of the two administrations by Group 1 were regarded as independent—as was argued by Wright (2003), for instance. Of course, the analysis is susceptible to maturation history or development factors with which the time span between the two test administrations affects the results of the latter test because of the development in ability. These two problems were unavoidable because we decided to use the

data for equating after administrations. That is, it was a post-hoc decision rather than all planned before test administration. As long as these issues remain, readers are suggested that, any claim based on the subsequent analyses of this study should be taken with caution.

Analysis

The following three steps comprise the main analysis of the present study. These steps follow the equating process suggested by Skaggs and Wolfe (2009).

Step 1: Separate analyses of Form A and Form B by using the Rasch model software, WINSTEPS (Linacre, 2006). In both analyses, person means were set to 0 with the standard deviation of 1 so that the results of the two analyses were comparable. In each analysis of Forms A and B, anchor items were used to calibrate measures using the test-taker responses derived from different versions of the test. This linking of data is called concurrent calibration.

Step 2: The item difficulty measures of anchor items generated in the first step were used to estimate a linear equation (a slope and constant). This linear equation predicts anchor item difficulties of Form B using those of Form A. A freeware called POLYST (Kim & Kolen, 2003) was used for the generation of the equation.

Step 3: The equating of the scores of the two Forms was implemented using a freeware called PIE (Hanson & Zeng, 2004). A linear equation generated in the second step was used to transform all item difficulties of Form B. Both the item difficulties of Form A and those of transformed ones of Form B served as input for the equating of scores of the two Forms.

Results

Table 3. Descriptive Statistics of Rasch analyses for Forms A and B

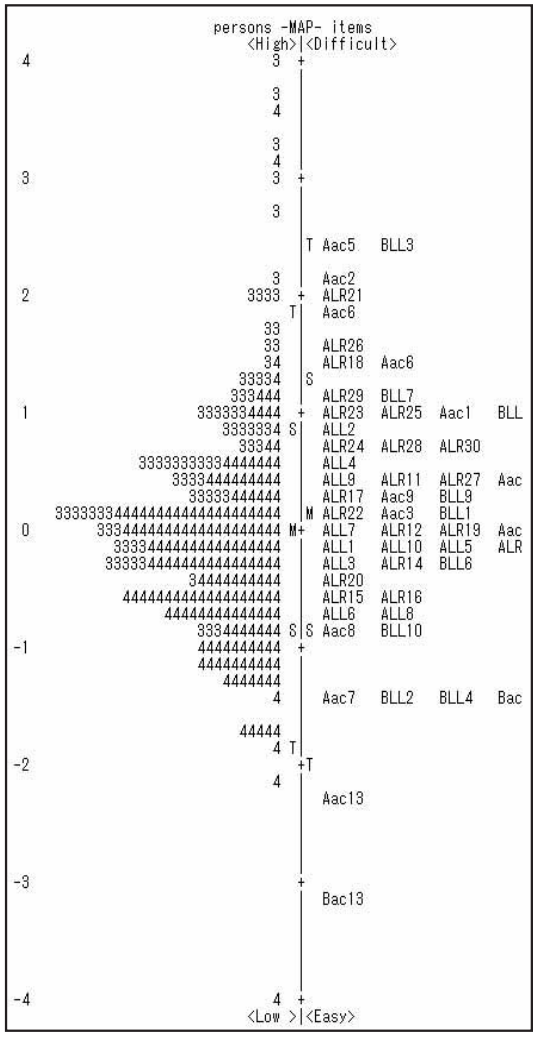
		Mean	SD	Separation	Reliability	Misfits
Form A	Person	.03	1.00	2.01	.80	0
	Item	.01	.91	4.89	.96	0
Form B	Person	.00	1.00	2.01	.80	0
	Item	-.03	1.38	5.39	.97	0

Table 4. Descriptive Statistics of Raw Total Scores for Forms A and B

	Mean	SD	Skewness	Kurtosis	*n*
Form A	18.77	6.17	.59	1.22	273
Form B	18.88	6.72	.05	.33	268

Step 1: Separate Rasch Analyses of Forms A and B

Table 3 shows the descriptive statistics of the results of the two Rasch analysis runs. The results suggest that both Forms have an acceptable reliability, and there were no misfitting items or persons if one applies the Fisher's fit and targeting criteria for "very good" measures (Fisher, 2007).



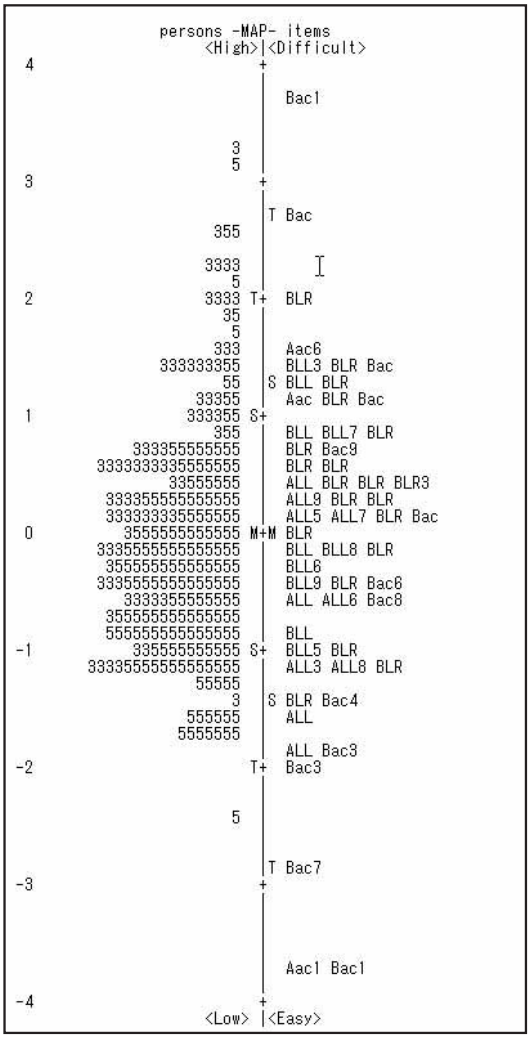


Figure 1 (left). The Wright map of Form A. Figure 2 (right). The Wright map of Form B. Both Figures show from the left logit scale, person ability measures plotted, and item difficulty measures. Person IDs 3, 4, 5 indicate data from the third (Groups 2 and 3), fourth (Group 1), and fifth (Group 1) administrations.

Table 3 shows the descriptive statistics of raw total scores. Both analyses suggest that the two Forms are fairly similar to each other. Figures 1 and 2 show the Wright maps of Forms A and B respectively. In both Figures, the left half shows the spread of person ability measures. Numbers 3, 4, and 5 indicate student IDs from the third administration (Group 2 and 3), and the fourth and fifth (Group 1) administrations. In both cases, test items seem to well target at most test-takers, and the range of item difficulties cover most test-takers' ability, with a few exceptions of top students in Form A, who were far beyond the coverage of item difficulties. Overall, these two Figures display a striking similarity between the two Forms.

Step 2: Estimating a Linear Equation of Anchor Items of Forms A and B

The 26 anchor item difficulty measures of Forms A and B were plotted, and there were 4 anchor items that displayed more than one logit difference between Forms A and B. Figure 3 shows the plot of anchor item difficulties after the four items, those that were off the diagonal, were eliminated.

Using POLYST, the remaining 22 anchor item difficulties of Form A were entered as input for estimating those of Form B. The analysis generated a linear relationship of $\beta_B = 1.03\beta_A + 0.09$, where

β_A is item difficulties of Form A, and β_B is that of Form B. The linear equation transformed the item difficulties of Form B, and those values were summarized in Table 5. Although the two Forms were, again, strikingly similar before scale linking, mean differences between the two Forms were reduced even further after scale linking.

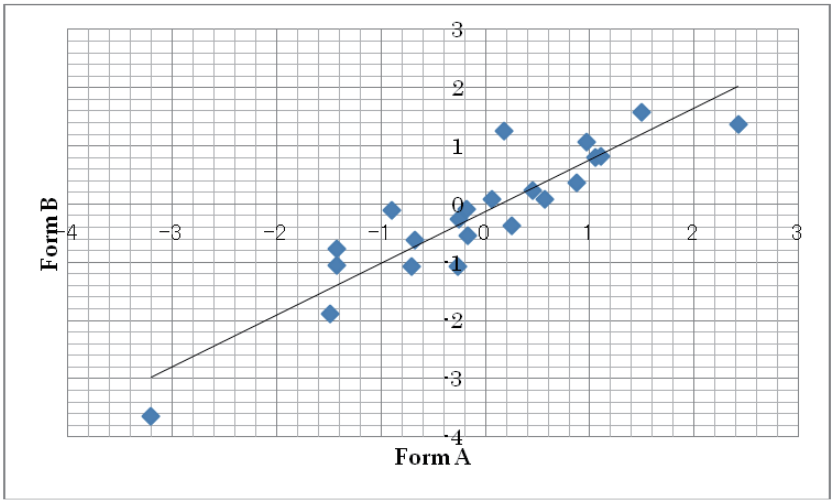


Figure 3. The figure shows the plot of anchor item difficulties of Form A against those of Form B after the four off-the-diagonal items were removed.

Table 5. Descriptive Statistics for Item Difficulties Before and After Scale Linking

	Before Scale Linking		After Scale Linking	
Form	Mean	SD	Mean	SD
A	.15	1.10	.15	1.10
B	.00	1.35	.25	1.15
A Anchor	-.05	1.21	-.05	1.21
B Anchor	-.17	1.18	.03	1.26

Step 3: The Equating of the Scores

Finally, the equating of the scores of the two Forms was implemented with item difficulties of Form A and the transformed item difficulties of Form B. Table 5 displays the results. The first column shows the raw scores of Form A. The second column shows the equivalent Rasch logit measures. The third column shows the equated score of Form B.

Conclusion

This study herein describes a process of equating of the two Forms of the listening prochievement tests for third-year junior high school students. Despite some drawbacks in the data setup, both raw data and separate Rasch analyses indicated that the two Forms are fairly similar, even without any equating. The successful linking and equating of the two Forms has brought us the two parallel forms of the listening prochievement test.

Table 6. Results of Score-Equating of Forms A and B

Form A Raw Score	Ability Measures	Form B Equivalent
0	-------	0.00
1	-3.26	1.02
2	-2.66	2.08
3	-2.29	3.15
4	-2.01	4.23
5	-1.79	5.31
6	-1.61	6.39
7	-1.45	7.46
8	-1.31	8.54
9	-1.18	9.61
10	-1.07	10.69
11	-0.96	11.76
12	-0.86	12.83
13	-0.76	13.90
14	-0.67	14.96
15	-0.58	16.03
16	-0.50	17.09
17	-0.42	18.14
18	-0.34	19.20
19	-0.27	20.25
20	-0.19	21.29
21	-0.12	22.33
22	-0.05	23.37
23	0.01	24.41
24	0.08	25.44
25	0.14	26.47
26	0.21	27.49
27	0.28	28.51
28	0.35	29.52
29	0.41	30.53
30	0.48	31.54
31	0.55	32.54
32	0.62	33.54
33	0.69	34.53
34	0.76	35.52
35	0.83	36.50
36	0.91	37.48
37	0.98	38.46
38	1.06	39.43
39	1.14	40.40
40	1.22	41.36
41	1.31	42.31
42	1.40	43.26
43	1.50	44.21
44	1.60	45.14
45	1.71	46.08
46	1.83	47.01
47	1.96	47.93
48	2.11	48.84
49	2.27	49.75
50	2.46	50.64
51	2.69	51.52
52	3.00	52.38
53	3.47	53.21
54	-------	54.00

Acknowledgement

This report is supported by the Grant-in-aid for Scientific Research awarded to the first author (no. 22520550). We also thank Minoru Akiyama for his great help.

References

Eiken Test in Practical English Proficiency Pre-2nd Grade (Eiken jyunnikyuu zenmondaisyuu CD) 2008. [CD]. Obunsha, Tokyo.

Fisher, W. P. 2007. Rating scale instrument quality criteria. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, **21**(1), 1095.

Hanson, B., and Zeng, L. 2004. PIE (computer program). The University of Iowa, Iowa City, IA.

Kim, S., and Kolen, M. J. 2003. POLYST (computer program). The University of Iowa, Iowa City, IA.

Linacre, J. M. 2006. WINSTEPS (computer program). Winsteps.com, Chicago.

Saito, H., Saito, T., and Kino, I. 2011. Development of listening prochievement tests for third-year Japanese junior high school students studying English as a foreign language (Part I). *茨城大学教育学部紀要（教育科学）*, **60**, 111-121. (http://ir.lib.ibaraki.ac.jp/bitstream/10109/2556/1/20110036.pdf)

Skaggs, G., and Wolfe, E. W. 2009. Equating designs and procedures used in Rasch scaling. Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models, E. V. J. Smith and G. E. Stone, eds., JAM Press, Maple Grove, MN, 364-383.

Tokyo Shoseki. 2006. New Horizon English Course 3: Teacher's Manual (Waakushiito hen 1), Tokyo Shoseki, Tokyo.

Total English 3 (Oyohatten ban waakushiito syuu). 2006. Gakko Tosho, Tokyo.

Wright, B. D. 2003. Rack and stack: Time 1 and Time 2. *Rasch Measurement: Transactions of the Rasch Measurement* SIG, **17**(1), 905-906.