研究論文

To what extent can self-assessment of language skills predict language proficiency of EFL learners in school context in Japan?

外国語能力自己評価項目は高校生・大学生の外国語能力をどの程度予測できるのか。

YOSHIZAWA Kiyomi
吉　澤　清　美

外国語能力自己評価項目（can-do項目）は学習者が自らの言語能力を評価するものであり、その価値については、学習者の勉学意欲を向上させる、自立した学習者の育成につながる、学習者のニーズ分析をする上で教師に多くのヒントを与える、外国語能力を測定する標準テストにかわるなどが指摘されている。Inoue (2008) は、英語圏での在住経験などがなく、外国語として英語を学ぶという日本の学習環境において、高校生・大学生が英語を使うことができる日常的なタスクを想定し、外国語能力自己評価項目を作成した。本研究では、Inoueの外国語能力自己評価項目の妥当性検証を行った。外国語能力自己評価項目は読解力に関する15項目、聴解力に関する10項目、計25項目から成り、それらを151名の大学入学時の学習者に実施し、ラッシュ測定モデルを用い、項目分析を行った。さらに、外国語能力自己評価項目がどの程度学習者の外国語能力を予測するかを検証した。

キーワード
can-do statements　self-assessment　foreign language proficiency
Rasch measurement models　cumulative response process

1. Introduction

Can-do statements are a form of self-assessment of foreign-language skills. Second or foreign language (L2) learners are asked to rate their abilities of performing tasks described in can-do statements. For example, I can read and understand a letter from my friend written in English. Self-assessment is also known as self-rating, self-appraisal, self-control, or self-evaluation.

Blanche and Merino (1989) summarized the literature on self-assessment of foreign language skills and pointed out that self-assessment accuracy would lead to learner autonomy and help teachers to become aware of learners' individual needs. Also, they reported that self-assessment practices "appeared to have increased the learners' motivation" (p. 324). Similarly, Ross (1998) conducted a meta-analysis on self-assessment of language skills and reported substantial correlations between various criterion measures and L2 learners' self-rating of their language skills. Ross argues for self-assessment as an alternative to "a more expensive and logistically viable approaches to proficiency and achievement assessment" (p.1). The present study examines the validity of newly developed can-do statements to assess reading and listening abilities of EFL learners in school context in Japan.

1.1 Self-assessment in second language testing

Blanche and Merino (1989) conducted an extensive literature review on self-assessment in language testing and presented a prose-based summary of self-assessment in language testing, including sample size, methodology and criterion variables to measure second and foreign language proficiency. One of the major findings they presented is a consistent overall agreement between self-assessments and ratings using various external criteria. They also included the studies which presented the quantitative comparisons between self-assessments and objective measures of proficiency. In those studies, Pearson product-moment correlation coefficients were often used and their values ranged from .50 or .60 to higher. On the other hand, Blanche and Merino reported two studies which showed no significant relationships between the accuracy of self-assessments of learners' language skills and their actual test (or classroom) performance.

Ross (1998) conducted a meta-analysis of self-assessment in the foreign and second language testing. He included the studies which empirically examined the relationship between self-assessment and four second language skill areas, namely reading, speaking, listening and writing, and made a summary of the meta-analysis of the 60 correlations. His summary suggests robust correlations between self-assessment and criterion skill measures. Further, he examined the effect of experiential factors in self-assessment. Beginning and elementary-level learners completed 20 skill-focused self-assessment items and a 60-item achievement test. The achievement items were designed to assess the skills and content covered in a coursebook used for a year-long English as a foreign language program. The format of a few of the test sections were modified though the functional content stayed the same. This manipulation was made to see whether self-assessment is most accurate when the criterion is based on

experiences learners had in classroom context or based on general proficiency. The results show that the self-assessment measure had considerably larger multiple correlations with the sections of the test which matched their classroom experiences than those with the modified format section. The finding suggests that "the episodic memory of using particular skills in the classroom experience would enhance the accuracy of self-assessment" (p. 16).

1.2 Can-do statements in relation to the Test of English for International Communication (TOEIC)

The TOEIC Can-Do Test was developed to provide information to help test users to interpret TOEIC scores. When a test taker receives a score in TOEIC reading comprehension or listening comprehension, what does the score mean? "Specifically, what can a person with such TOEIC scores actually do in a business setting with English?" (The Chauncey Group International, 2005, p. 38) To this end, the Can Do Research Study was conducted in 1995 by Research Division of Educational Testing Service and the International Institute of Business Communication of Japan.

The research group selected 75 can-do statements from the previous research studies which dealt with self-assessment of language abilities. Those can-do statements:

(1) described concrete tasks;
(2) described tasks likely to be familiar to TOEIC test-takers;
(3) described tasks related to work settings;
(4) described tasks likely to be meaningful to those who interpret and use TOEIC scores; and
(5) reflected both the business and the social aspects of work (*TOEIC Can-Do Guide*, p. 4).

8,601 Japanese TOEIC test-takers were asked to rate their abilities to perform tasks in a business setting using English in the five performance areas: reading, writing, speaking, listening, and interactive skills. These self-ratings were matched with their TOEIC scores to make correspondence tables. The TOEIC reading scores are divided into five groups and a correspondence table is created for each group. Each correspondence table describes the tasks test-takers can do, those they can do with difficulty, and those they cannot do in the performance areas of reading and writing. The same procedure was repeated for the correspondence tables between the TOEIC listening scores and the tasks in the performance areas of listening, speaking, and interacting.

Powers, Kim, and Weng (2008) administered a self-assessment inventory to TOEIC test

takers in Japan and Korea to obtain their perceived abilities to perform various reading and listening tasks in everyday life. Approximately 10,000 examinees took the redesigned TOEIC (reading and listening) and answered 50 can-do statements. The researchers found the correlation results were congruent with those reported in validity studies using different kinds of validation criteria such as course grades and supervisors' ratings.

1.3 The purpose of the present study

More and more schools have started to use TOEIC in formal school settings in Japan. The International Institute of Business Communication of Japan conducted a survey on the use of TOEIC in September and October, 2007. 1,769 institutions responded to the survey, including both graduate and undergraduate programs. The survey reported that those institutions used TOEIC for admission, placement, or providing credits (The International Institute of Business Communication of Japan, 2008).

In spite of the fact that TOEIC has been used widely in school settings, it is less likely that the TOEIC Can-Do statements can be applied to EFL learners at school settings without adapting the tasks in some of the statements. This is due to the aforementioned third and fifth criteria used to select can-do statements from the previous research studies which dealt with self-assessment of language abilities: (3) the can-do statements "described tasks related to work settings" ; (5) the can-do statements "reflected both the business and the social aspects of work." If TOEIC is used for admission or placement in a study program, it is more likely that test-takers are high-school or university students and they are not familiar with tasks related to work settings or business aspects of work. TOEIC users, i.e., test takers, school administrators, and instructors, would like to relate the test scores on TOEIC reports to the tasks test takers can perform or perform with difficulty. At the moment, there is little study conducted to examine the relationship between TOEIC scores and the performance level which test takers perceive that they can carry out using English in non-business settings.

The purpose of the present study is twofold. First, the study aims to examine the validity of can-do reading and listening statements developed in Inoue (2008). He thoroughly examined the TOEIC can-do statements and developed new reading and listening can-do statements to assess EFL learners' performance of everyday language tasks, i.e., reading and listening tasks, in English. The tasks in his can-do statements were originated from his one-year observation of EFL learners in the school context in the western part of Japan. The targeted group is EFL learners at the entry level to a language program at a university or at an equivalent level in Japanese school context. The detailed description of the can-do statements is presented in the

Method section below. Specifically, the present study examines the following two aspects of the can-do statements: the relevance of tasks in the can-do reading and listening statements; the appropriate number of response categories. The second purpose is to examine to what extent the developed can-do statements predict reading and listening abilities of EFL test takers in Japanese school context.

2. Methods

2.1 Instruments

Can-Do statements to perform everyday language tasks in English. To develop can-do statements, Inoue used the following guideline. Can-do statements:

（1）describe concrete tasks;

（2）describe tasks familiar to a targeted group in the EFL context in Japan;

（3）describe the tasks which vary conceptually in terms of the amount of language processing（reading and listening materials）.

The first point is based on the findings of Blanche and Merino and those of Ross. They both indicated that the accuracy of self-assessment items would increase when the items contain "the descriptions of concrete linguistic situations"（Blanche & Merino, p. 324）. The second point is based on Ross who provided the evidence to show the importance of an experience factor in self-assessment. The last point is related to a self-assessment items which reflect a cumulative response process. That is, if respondents have more of the construct of interest, i.e. reading and listening abilities, they would respond more positively to the tasks which require more language processing.

The developed can-do statements differ from the TOEIC can-do statements in two respects. First, the tasks in the new can-do statements are targeted for EFL learners in Japanese school context. Table 1 presents the tasks in the new can-do statements. The tasks in the developed can-do statements were targeted at the EFL learners in the upper classes at a senior high school and lower classes at a university level. Based on the tasks in Table 1, 15 can-do reading statements and 10 can-do listening statements were developed. The second difference is that the new can-do statements have six response categories, whereas the TOEIC can-do statements have five. This change is made to avoid a situation where respondents prefer to take a neutral position, that is, 3 in a five-point scale.

Table 1　The tasks included in the developed can-do statements

Reading Tasks
• memos like a shopping list
• storefront signs
• table of contents in an English book
• restaurant menus
• topics based on headlines of newspaper articles
• the content based on the title of a book
• signs and information in a bus or train
• instructions or explanation
• self-introduction written in English
• stories and conversations in an English textbook
• lyrics in English songs
• a brochure for a study-abroad program
• an English newspaper
• information on the internet
• novels or stories in English
Listening Tasks
• announcement in a bus or train, including its destination, departure time, and arrival time
• self-introduction spoken in English
• a Japanese teacher speaking in English in class
• Japanese animations dubbed in English
• a movie with subscripts in English
• a movie without subscripts
• a foreigner speaking to a respondent in English
• the content of a radio program or information provided in the program
• lyrics in English songs
• discussion conducted in English

Practice Reading and listening Tests. Participants took TOEIC practice reading and listening tests and their results were used as measures of their reading and listening abilities.

2.2 Participants

151 university students participated in the study. They were freshmen at a four-year university in the western part of Japan. 90 of them majored in economics and 61 majored in commerce. All the participants answered the developed can-do statements. One half of the participants took the practice reading test; the other half took the listening part.

2.3 Data Analyses

The responses to the developed can-do statements were analyzed using WINSTEPS (Linacre & Wright, 2000) and RUMM2020 (Andrich, D., Sheridan, B., & Luo, G., 2003), Rasch unidimensional measurement model software. The Rasch model is the only probabilistic model, which provides "the necessary objectivity for the construction of a scale" (Bond & Fox, p.7)

and item difficulties are calibrated independently of the attributes of the people who take them. The application of these programs requires that each item on a test or a questionnaire contributes to the measure of a single trait. Prior to the analyses based on the Rasch model, principal component analyses were conducted to confirm unidimensionality. The results indicated that there is one large component and each of the can-do statements contributes to the measure of a single trait.

The can-do test data was first examined to the extent that the tasks in the can-do statements would reflect cumulative response process. Next, threshold analysis was conducted to see whether each response category functions as it should. Similarly, the responses to the reading and listening practice tests were analyzed. At the last step, person measures were calculated based on the analyses of can-do statements and the practice reading and listening tests, and the correlations between the self-assessment measures and the reading and listening measures were examined.

3. Results and Discussion

3.1 The number of response categories

The data was first analyzed by WINSTEPS to examine the adequacy of the number of response categories. They were analyzed first with six response categories; then, categories five and six were merged as one response category. Figures 1 and 2 show the item maps, distributions of items and persons on the common scale, which is a dotted line in the middle. The persons are shown on the left; the items are shown on the right. M on the common scale refers to means; S, one standard deviation; T, two standard deviations. The higher the items are, the more difficult they are. Similarly, the higher the locations of the persons are, the more proficient they are. Figure 1 shows the item map with six response categories; Figure 2 shows the item map with five response categories. These figures clearly indicate that six response categories make the can-do statements more difficult than the five response categories. Further, there are not enough items for people with lower abilities. Also, it is assumed that a person who chooses higher categories would have mean abilities higher than those who would choose lower categories. However, 12 items showed the mean abilities are not ordered when six response categories were used. Based on these results, further analyses were conducted with five response categories.

The first two columns of Table 2 on page 74 show the summary of the fit statistics for the can-do statements with 25 items. When the data fit the measurement model, the fit statistics

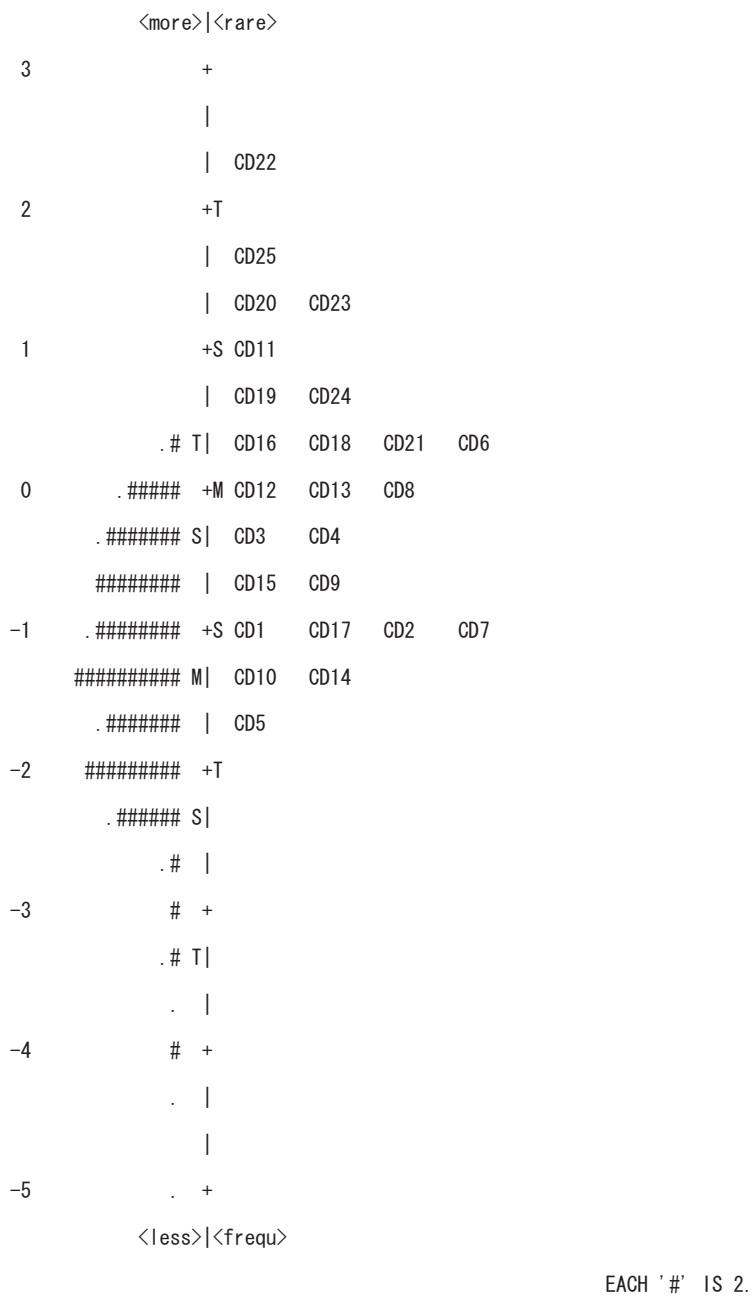

Figure 1　Item Map (6 categories)

have a mean near zero and standard deviation near one. The summary shows that the data fit the model. Also, the item-trait interaction *chi-square value* is 62.714, *df* = 50, *p*= .107, showing there is no significant interaction between the responses to the items and the location

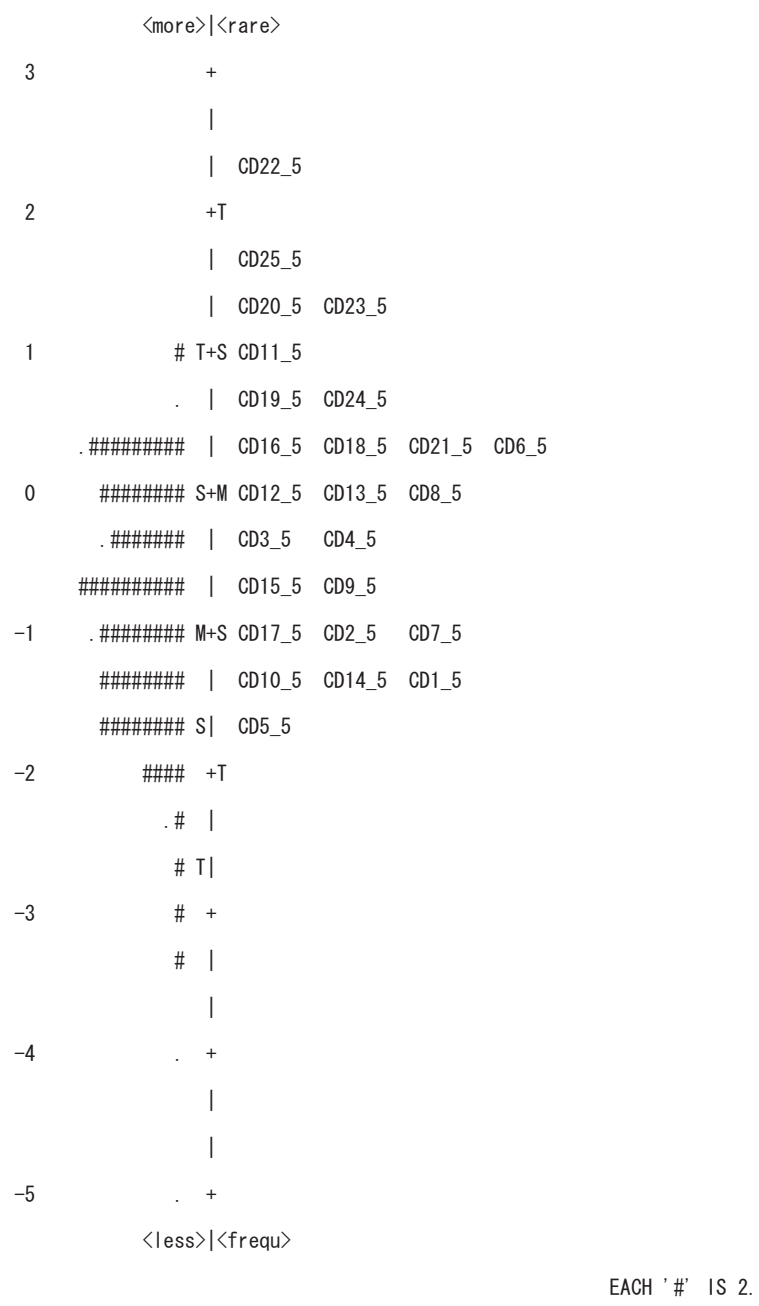

Figure 2 Item Map (5 categories)

values of the students along the trait. This indicates that the dominant trait is affecting all the responses to the items and there is reasonable agreement about the difficulties of the items. The person separation index, an equivalent of reliability in Classical Test Theory, is 0.92,

Table 2　Fit statistics for the can-do statements

	25 Items		21 Items	
	Items	Persons	Items	Persons
Location mean	0.000	−0.926	0.000	−0.819
SD	1.231	0.929	1.190	0.965
Fit statistic mean	0.303	−0.220	0.265	−0.246
Fit statistic SD	1.037	1.568	0.922	1.534

indicating the person measures are well separated in relation to the measurement errors.

3.2 Threshold analyses

Bond and Fox define a threshold as “the level at which the likelihood of failure to agree with or endorse a given response category (below the threshold) turns to the likelihood of agreeing with or endorsing the category (above the threshold)” (p. 234). That is, a threshold is a point at which the probability of selecting two adjacent categories is the same. With polytomous data, including the data on Likert scale, thresholds must be ordered. If thresholds are disordered, scoring is not functioning as expected. Table 3 shows the results of threshold analyses. The four asterisks on the right hand of the table indicate that the thresholds of those items are disordered: CD13, 21, 22, and 24. CD 13, 22 and 24 show that the threshold between response categories 4 and 5 (threshold 4) is lower than the threshold between categories 3 and 4 (threshold 3); CD 21 show that the threshold between response categories 4 and 5 (threshold 4) is lower than that between categories 2 and 3 (threshold 2). This means that a person needs more ability to choose between 2 and 3 than the ability needed to choose between 4 and 5. This is counterintuitive. Tasks described in can-do statements with disordered thresholds are:

CD 13　Japanese animations dubbed in English

CD 21　Watching a movie with subscripts

CD 22　Watching a movie without subscripts

CD 24　Novels or stories in English

Figure 3 shows the category characteristic curve for CD13 and illustrates the disordered thresholds visually. The x axis shows person location (ability) in logits. The y axis shows the probability of a person with a given ability would respond correctly to the item with a certain level of difficulty. Figure 3 shows a curve with 0 on the left-hand side. This indicates that as the ability of a person increases, the probability of a score of 0 decreases. Similarly, there is a curve with 4 on the right. This shows that as the ability increases, the probability of a maximum score increases. Three curves between these two curves indicate the following.

Table 3 Thresholds

Item		Thresholds				
Code	Location	1	2	3	4	
CD1	1.478	−2.852	−0.371	1.136	2.087	
CD2	1.118	−2.368	−0.799	1.138	2.029	
CD3	−0.377	−1.927	−0.772	0.673	2.027	
CD4	−0.393	−2.356	−0.951	1.086	2.221	
CD5	1.628	−1.706	−0.807	0.568	1.945	
CD6	0.394	−2.166	−0.714	0.966	1.914	
CD7	1.039	−2.883	−0.336	0.749	2.470	
CD8	−0.127	−2.018	−0.211	0.794	1.435	
CD9	−0.670	−2.103	−0.595	0.568	2.130	
CD10	1.518	−3.235	−0.257	0.718	2.774	
CD11	0.257	−0.898	−0.109	0.499	0.508	
CD12	−0.135	−1.784	−0.567	0.165	2.186	
CD13	−0.311	−1.473	−0.692	1.360	0.805	*
CD14	1.563	−2.520	−0.746	0.926	2.340	
CD15	−0.451	−2.268	−1.284	0.446	3.106	
CD16	0.039	−2.541	−0.262	1.123	1.680	
CD17	−0.974	−1.277	−0.608	0.345	1.539	
CD18	1.438	−3.248	−1.883	−0.089	5.220	
CD19	1.559	−3.025	−1.958	−0.500	5.483	
CD20	1.724	−2.733	−0.790	0.083	3.440	
CD21	−0.273	−1.052	−0.003	1.560	−0.505	*
CD22	2.722	−2.244	−2.113	4.638	−0.281	*
CD23	1.825	−2.961	−1.217	0.122	4.057	
CD24	0.126	−1.822	−0.427	2.244	0.004	*
CD25	1.972	−2.108	−1.225	−0.133	3.465	

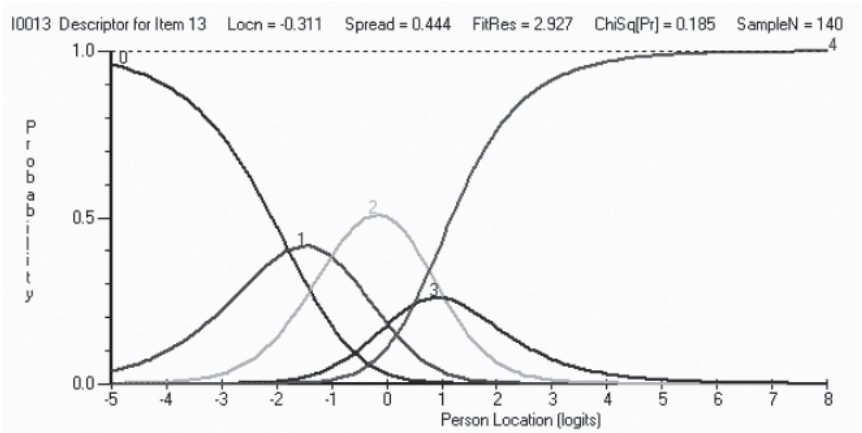

Figure 3 The category characteristic curve for CD13

Concerning Curvel, when a person has a very low ability relative to the item's difficulty, the probability of a response of 0 is most likely. When a person has ability much higher than the item's difficulty, then the most likely response is 2. When a person is of moderate ability relative to the item's difficulty, the most likely response is 1. The same is true with curves 2 and 3 although the most likely responses are 1, 2, and 3 for curve 2; 2, 3, and 4 for curve 3. Figure 3 shows that curve 3 is lower than other curves and does not indicate any specific point on the person ability continuum, x-axis. Also, curves 3 and 4 meets before curves 2 and 3 meets, indicating the disordered thresholds.

The four items with disordered thresholds were deleted and threshold analysis was repeated. The two right columns on Table 2 on page 74 present the summary of the fit statistics for the can-do test with 21 items. The summary shows that the data fit the model in general and the item fit statistics improved. Also, the item-trait interaction *chi-square value* is 47.064, df=42, p= .273, showing there is no significant interaction between the responses to the items and the location values of the students along the trait. The separation index is 0.908.

Figures 4 to 6 show the threshold maps: Figure 4 with 21 items, Figure 5 with fourteen reading items and Figure 6 with seven listening items. Items on the left column are ordered in terms of their difficulties. These three figures indicate several findings. First, the thresholds between Category 1 and Category 2, shown as 0 and 1, indicate that thresholds are gradually becoming higher as the difficulty level of the items increases. On the other hand, the distances between thresholds are not evenly divided within an item. Especially, this phenomenon is

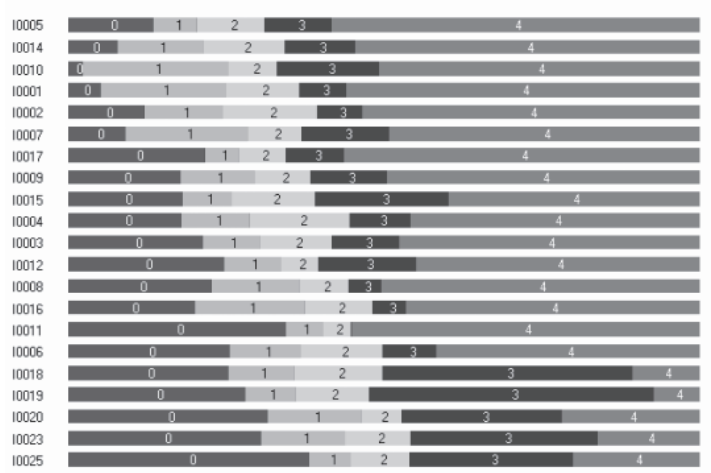

Figure 4　Threshold map with 21 items

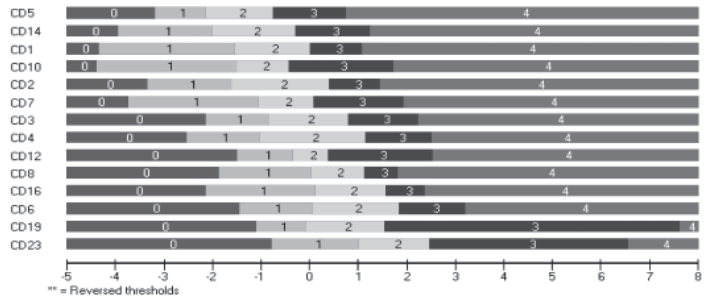

Figure 5　Threshold map with reading items

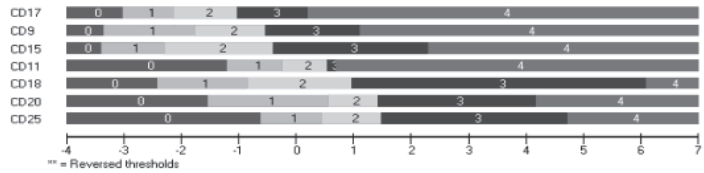

Figure 6　Threshold map with listening items

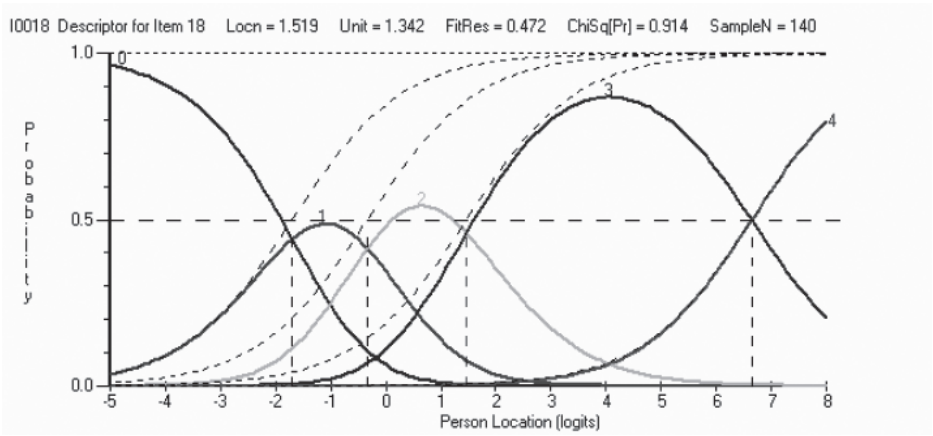

Figure 7　Category probability curves of item 18

observed clearly in most difficult five items, items 18, 19, 20, 23, and 25. To illustrate this point, Figure 7 shows the category probability curves of item 18. The overlaying dotted lines indicate threshold probability curves. The vertical dotted line meets with the x axis (Person Location), persons' ability continuum. The first three thresholds fall on between minus 2 and plus 2; the last threshold, threshold 4, falls further right on the continuum, indicating a person has to be extremely capable to choose category 5 of item 18. These last five items are too difficult for the target population, and they may need to be replaced in a revised version.

Actually, the first three categories of these five items were selected by 90% or more respondents and less than 10% selected category 4 and none selected category 5. Similarly, very few respondents chose the last two categories of item 11; more than 90 % of them chose categories 1 to 3. Although the thresholds are ordered, there is very little ability difference between the respondents who chose category 4 and those that chose category 5, 0.478 and 0.498 in logits.

3.3 Targeting

Figures 8 and 9 plot the person-item threshold distributions with 25 and 21 items,

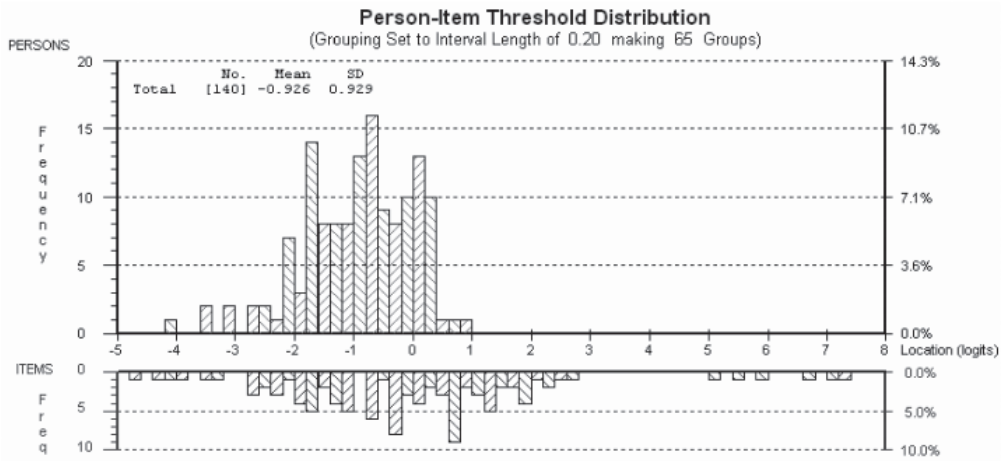

Figure 8　Person-Item Threshold Distribution (25 items)

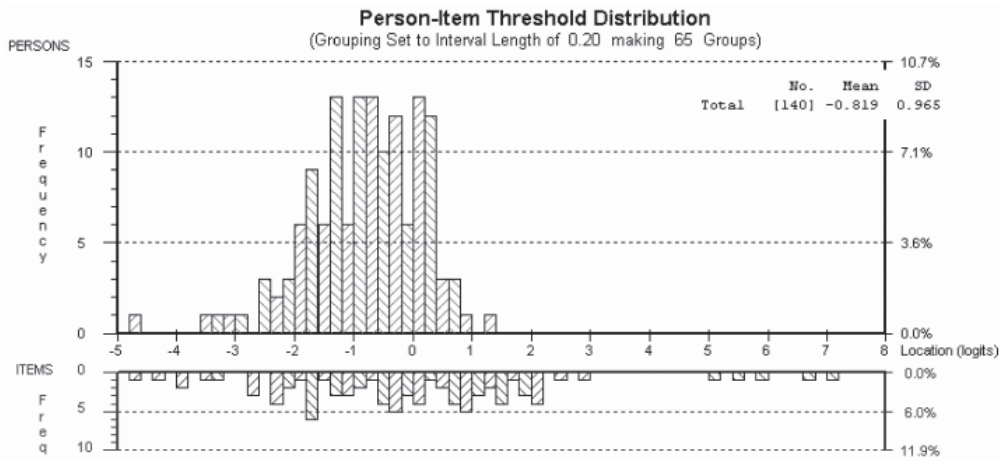

Figure 9　Person-Item Threshold Distribution (21 items)

respectively. The top graph shows the distribution of persons and the bottom shows that of items on the common scale. As presented in Table 1, the mean of items is zero, but the mean of the persons is between zero and minus one. This indicates that the items are a little difficult for the target group. Also, the standard deviations of the items are larger than those of persons, indicating items are more spread out than persons are. Also, both figures show the items to the farther right are off-target.

3.4 Correlations between the can-do measures and language-skill measures

Table 4 presents the correlations between the can-do person measures and language-skill measures. The correlations between can-do abilities and language-skill measures are significant when all the twenty-one items were included as a proficiency measure (r = .358 with the TOEIC reading items, r = .26 with the TOEIC listening items). Similarly, when only can-do reading items were included, the can-do reading measure showed significant correlations with language-skill measures (r = .388 with the TOEIC reading items, r = .307 with the TOEIC listening items). Also, the can-do reading measure showed a higher correlation with the reading-skill measure than with the listening-skill measure. On the other hand, can-do listening measure did not show a significant correlation with the listening-skill measure. This presents some evidence that the can-do reading items have concurrent validity though its correlation with the criterion measure is low. On the other hand, the correlation study did not present any support for the can-do listening items in terms of validity.

Table 4 Correlations among the can-do measures and language-skill measures

	Language-Skill Measures	
	Reading	Listening
Can-Do reading	.388**	.307**
Can-Do listening	.178	.122
Can-Do all	.358**	.260*

**p=0.001 *p=0.025

The findings of the present study are congruent with those of previous studies which used TOEIC test. Powers et al. administered two forms of can-do statements to TOEIC test takers in Japan and Korea to obtain their perceived abilities to perform various reading and listening tasks in everyday life. The correlations between can-do listening statements and TOEIC listening scores were .53 for two forms; those between can-do reading statements and TOEIC reading scores were .46 to .47 respectively. The correlation coefficients found in the present study is not as high as those in Powers et al. However, we can say that the correlation

coefficient between the can-do reading statements and the reading-skill measure in the present study is approaching those in Powers et al. About 15% of the variance in the reading test and the can-do reading statements are shared.

Why is the correlation between the can-do reading statement and the reading-skill measure lower in the present study than the previous studies? Blanche and Merino observed a consistent overall agreement between self-assessments in the studies they examined, though they also detected considerable variations in the accuracy of students' self-assessment: "The accuracy of most students' self-estimates often varies depending on the linguistic skills and materials involved in the evaluation" (p. 315). They state that the studies including self-assessment and objective measures of proficiency often report the Pearson product-moment correlation coefficients ranging from .5 to .6 and "higher ones are not uncommon" (p. 315). Similarly, Ross reports that the average correlation between self-assessment and the criterion variables for reading skill is .61. One of the factors contributing to a low correlation between the self-assessment and the reading-skill measure in the present study may be the discrepancy between the tasks described in the can-do statements and the abilities of the test-takers. As the item map (Figure 2) shows, the item mean (M on the right side) is set at zero; on the other hand, the mean of the persons (M on the left side) falls around minus one. That is, the items are more difficult than the abilities of the test-takers. Also, the item map shows there is no item whose difficulty level matches less proficient learners, i.e., the items further down the common scale. The person-item threshold distribution with 21 items (Figure 9) also shows that the items on the far right do not have persons whose abilities match their difficulty levels. Another factor may be the ability distribution of the participants of the present study. Since the intact groups were used for the present study, there was no guarantee that they were representative of the population and it is likely that the variance of their abilities might be limited.

Also, the present study shows an insignificant correlation between the can-do listening statements and the listening-skill measure. One of the factors may be the limited number of listening can-do statements. There were ten items, but three items were deleted. The tasks in seven items are not enough to describe what EFL learners across different proficiency levels can perform. Another factor may be that there is a mismatch between the experience the participants in the present study had with listening and the tasks described in the can-do statements. Based on the meta-analyses, Ross shows that self-assessment for reading skill "is relatively more valid than that of lesser developed skills" (p.6). In his study, the self-assessment of listening skills shows a strong average correlation, .65. However, there is a

wider range of variation in the learners' accuracy in the self-assessment of listening skill than the reading skill. Ross lists several factors which may affect self-assessment. One of the factors is called experience factor, that is, whether the tasks in the self-assessment items are directly related to the second language learners' experience with the language either through instruction or language contact. Ross states that the listening experience which learners in EFL context have is less extensive than their experience with reading. If learners have limited experience with the second or foreign language and their responses to the can-do statements may result in a method artifact. A similar thing might have happened to the participants in the present study in the process of assessing their listening ability.

4. Conclusion

The present study examined the validity of can-do reading and listening statements, which were developed to assess EFL learners' performance of everyday language (reading and listening) tasks in English. One of the aims is to examine the relevance of tasks in the reading and listening can-do statements. 21 out of 25 items showed that their response categories were ordered. The analyses of the data indicate that those remaining 21 items reflect cumulative response process to a certain extent. Also, the can-do statements are spread along the ability continuum of respondents, though the spread of items is larger than the spread of people. Targeting should be improved. Also, the present study found that six response categories make can-do statements more difficult than five response categories.

The second aim of the study is to examine to what extent the developed can-do statements predict reading and listening abilities of EFL test-takers in Japanese school context. The results indicate that one unit change in the can-do statements leads to .388 unit change in the reading ability, but no significant correlation was observed between the can-do listening statements and the measure of listening ability. The results of the present study indicate a positive relationship between the can-do reading measure and reading-skill measure. As Blanche and Merino (1989) pointed out, self-assessment accuracy would lead to learner autonomy and help teachers to become aware of learners' individual needs. If self-assessment practice is also helpful to raise learners' awareness about their abilities and other aspects of learning, further studies with revised and additional can-do statements will be worth investigating for both learners and teachers in school context similar to those in this study.

The present study is not without limitations. The major limitation is that the participants are not randomly selected and generalizability of this study is limited. Another limitation is

the small number of can-do listening statements. Three deleted statements were related to listening tasks and only seven out of ten listening can-do statements functioned as expected. Those deleted tasks refer to watching movies and Japanese animation dubbed into English. In retrospect, it is less likely that learners watch foreign movies or animation in English when subtitles or dubbing is readily available in Japanese. It deserves further research on the listening tasks which EFL learners are likely to experience in their language learning context.

The educational implication of the present study is that can-do statements can be developed at an in-house language program in school context. In-house can-do statements can be based on the content of lessons or course objectives. They can be used as a tool to facilitate learner autonomy and as an alternative form for instructors to assess language proficiency of their students and to analyze their needs.

Acknowledgement

The author is very grateful to Mr. Ryo Inoue for providing the can-do statement data for the present study. She is also very thankful to the students who participated in this study.

References

Andrich, D., Sheridan, B., & Luo, G. (2005). *Rasch Unidimensional Measurement Models: A windows-based item analysis employing Rasch models* [Computer software]. Perth: RUMM Laboratory.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning, 39* (3), 313-338.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

The Chauncey Group International. (2005). *Technical Manual for the Test of English for International Communication.* Princeton, NJ: ETS.

ETS. (n.d.) *TOEIC can-do guide: Linking TOEIC scores to activities performed in English.* Princeton, NJ: ETS.

ETS. (2008). *TOEIC tesuto nyugakushiken, taninninteiniokeru katsuyoujoukyou* [TOEIC test: how TOEIC is used for admission and provision of credits]. Princeton, NJ: ETS.

Inoue, R. (2008). [The reading and listening can-do statements for EFL learners in school context in Japan: a pilot test]. Unpublished raw data.

Linacre, M., & Wright, B. (2000). *WINSTEPS* [Computer software]. Chicago: MESA Press.

Powers, D., Kim, H., & Weng, V. (2008). *The redesigned TOEIC (Listening and Reading) Test: Relations to test-taker perceptions of proficiency in English* (ETS Research Report No. RR-08-56). Princeton, NJ: ETS.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1-20.