Research Paper

A Pilot EFL Oral Communication Test for Japanese Elementary School Students

日本人小学生を対象としたオーラルコミュニケーションテストの試験的開発

MARTIN Ron

Key words: 小学校、英語教育、英語能力、評価基準、外国語活動

elementary school, English education, English language proficiency, evaluation standard, foreign language activities

Abstract

This paper describes a pilot research project about the development of an oral communication proficiency test for Japanese public upper elementary school English as a foreign language (EFL) learners. Foreign language assessment of elementary school children is a new, but a quickly expanding area of L2 research worldwide (McKay, 2006). This study confronts the issues of the EFL learning context, the lack of language assessment training of teachers, and what proficiency means in the elementary EFL context. This study also provides an oral communication proficiency standard in the form of a 4-point, three domain rubric covering communication competence, vocabulary/syntax and interactional competence. Many-facet Rasch Measurement analysis is applied to the three facets of this study: (a) 4th-grade elementary school students (*N* = 36), (b) the three rubric domains and (c) three raters.

1. Introduction

There is a growing seriousness about the potentials and pitfalls of English as a foreign language (EFL) elementary school programs (Cameron, 2001, 2003; Curtain & Dahlberg, 2004; Munoz, 2006; Nikolov & Curtain, 2000; Nikolov & Djigunovic, 2006). A majority of such programs have been implemented based upon the lay theory that younger is better regarding language proficiency (Nikolov & Curtain, 2000), but studies have yet to verify this contention (Munoz, 2006; Nikolov, 2009). It follows that foreign language assessment of elementary school children is a quickly expanding area of L2 research worldwide (McKay, 2006), and programs hope to show gains in language achievement or at the very least show that elementary school language education is worthwhile (Johnstone, 2000). This paper describes the development, implementation and results of a pilot oral communication proficiency test for Japanese public upper elementary school EFL learners.

2. Literature Review

2. 1. Language assessment

Though it is beyond the scope of this paper to address the history of language assessment, it must be noted that the seminal work by Canale and Swain (1980) and Bachman and Palmer (1996) heavily influenced this study.

Canale and Swain (1980) made the distinction between communicative competence and performance. They stressed that while paper-and-pencil testing may be able to show a learner's knowledge (communicative competence), such testing could not show a learner's ability to use language in context (performance). Bachman and Palmer (1996) further developed the idea of performance in language assessment by highlighting the need for task and language authenticity as well as the need for interaction between the test taker and the task itself. Thus, with the understanding of the need to assess performance and of language performance itself, test developers set out to create alternatives to traditional language assessment designs that would put the learner in an active and interactive role. Task-based language teaching (e.g., Long & Porter, 1985; Nunan, 1989) provided this alternative approach.

The use of task-like performance-based assessments was promoted by Norris, Brown, Hudson and Yoshioka (1998), but their use was also heavily critiqued by Brown and Hudson (1998) who voiced a strong opinion about the need for reliable and valid assessment designs, regardless of their alternative status. McKay (2006) underscored these concerns in addition to a range of other concerns related specifically to young learners.

2. 2. Concerns related to assessing young learners

There are primarily three areas related to assessing young learners that must be addressed other than test reliability and validity. These areas are the context of the language program, who does the assessment and what does proficiency mean when applied to young learners.

2. 2. 1. Context of the language program

The most common young language learner programs are the following types of language education: bilingual, immersion, English as a second language (ESL), and EFL. Yet even within each of these types of language programs, it is not uncommon to see differences within and among school districts which create problems for assessment. Johnstone (2000) listed a number of such problems which include, but are not limited to, the starting age of learners, the number and length of lessons, teachers' abilities, and the place of language learning within the overall curriculum.

2. 2. 2. Who does the assessment

Elementary school teachers lack the knowledge, ability and training to develop and administer EFL assessments (Johnstone, 2000; McKay, 2006; Nikolov, 2000; Rea-Dickins, 2000). Formal language teaching commonly begins at the secondary grade level, and thus, no formal training about language assessment has been provided to elementary school educators (Johnstone, 2000; McKay, 2006). Johnstone (2000) predicted, correctly so, that this lack of language assessment training would lead to the need to have outside organizations and researchers to develop assessment designs.

2. 2. 3. The meaning of proficiency

Currently, many elementary school programs worldwide lack coherent curriculum guidelines for children's foreign language study (Nikolov & Curtain, 2000). In addition, due to the variety of and the variation within elementary school language programs and because elementary school teachers do not have training in language assessment, deciding what language proficiency means with regard to elementary school language learners has become the job of outside organizations and researchers. Perhaps the most influential organization at this time is the Council of Europe. In 2001, the Council of Europe published the Common European Framework for Reference (CEFR) for the multilingual European context and has since published a number of other articles and studies on the assessment of young language learners. The CEFR utilizes the creation of

'Can-do' statements made by the Association of Language Testers in Europe, which separates students across six levels and the three domains of speaking/listening, reading and writing (Council of Europe, 2001). Hasselgreen (2000) reviewed the use of the CEFR and concluded that while the 'Can-do' statements seemed to define different language levels of young language learners appropriately, it did not adequately cover all types of student performance, nor could it be used as a stand-alone tool to determine a learner's level with confidence.

In Japan, the nationally recognized English language proficiency test is published by the Society for Testing English Proficiency ([STEP], n. d.) and is known as the STEP test. Dunlea and Matsudaira (2009) aimed to match levels of the STEP test to the CEFR. The investigation was not an attempt to link the two tests empirically, but rather to enable educators to talk about the STEP test to audiences outside of Japan. Dunlea and Matsudaira (2009) asked 10 judges who were familiar with the STEP test to independently interpret and rate minimal level expectancies between the upper levels of the STEP tests and the upper levels of the CEFR. Results showed low inter-rater reliability, and thus, no common equivalency yet exists. However, because STEP tests and the CEFR were made for different populations of language learners, the validity of such an investigation is questionable and without actual participant scores on the two tests, it would seem that independent raters who only attempt to equate two rubrics is not a reliable method, even if the goal was only to aid in communication among educators.

In sum, the CEFR is not considered to be a stand-alone tool to assess young EFL learners' language proficiency (Hasselgreen, 2000), and there is no link, empirical or theoretical, between the CEFR and the STEP tests. Therefore, there is no international or national oral communication standard for elementary school children that is applicable to Japanese children.

2. 3. Language assessment of young learners in Japan

Japan also faces the same concerns related to assessing young learners. Even though the Ministry of Education, Culture, Sports, Science and Technology (MEXT) announced that compulsory foreign language education at the elementary school level will begin in April 2011 for 5th and 6th-grade students (2008), a number of school districts have been providing English language lessons for a number of years, albeit under a variety of conditions regarding number of lessons, grade levels taught and who the primary teacher was (Butler, 2007). Moreover, language education at the elementary school level in Japan is decentralized, which puts all decision making responsibility on public school boards and each individual school allowing for even greater diversity between programs, and thus, program outcomes (Butler, 2007). Furthermore, elementary EFL activities are not viewed as language study, but rather exposure to and experience in communication in a foreign language (MEXT, 2008). Therefore, a policy level definition of language proficiency for elementary school learners does not exist. Such conditions make the assessment of language at the program level difficult, and on the national level, problematic.

In addition, Japan views secondary school as the beginning of formal language education (MEXT, 2003) and the teacher training of elementary school teachers has focused on the language proficiency of teachers as well as the team-teaching of language lessons with native English speakers (Butler 2004, 2007; MEXT, 2001). No reference has been made in the literature with regard to the training of Japanese elementary school teachers on language assessment. Moreover, MEXT (2002) referred to levels within the STEP test series as benchmarks for both lower and upper secondary school students, and in doing so, gave credence to such tests in general. In 2004 the Jr. STEP test, focusing on young children's aural language proficiency, was introduced (STEP, n.d.). Butler and Takeuchi (2006, as cited in Katsuyama, Nishigaki & Wang, 2008) administered the Jr. STEP Bronze test (STEP, n.d.) and found that 5th and 6th-grade public elementary school students did relatively well. Thus, the assessment of Japanese elementary school students outside organizations and researchers has already begun.

In summary, the Japanese elementary language education system is decentralized and elementary EFL activities are not viewed as language study. Japanese elementary school teachers have not been trained in language assessment, nor, under current policy decisions, will they need to be. Thus, the definition of language proficiency, and the investigation of it, are being left to outside agencies.

3. Oral Communication Proficiency Test Development Project

3. 1. Stakeholders

3. 1. 1. The principal

This project originated with an elementary school principal, who began his post in April 2010. He said he wanted to know about the students' ability to use English. He hoped that test outcomes would highlight areas for curricular reform, teacher training needs and positive student feedback to bolster the students' feeling of success. His focus was to view any outcome in a positive, yet actionable manner.

3. 1. 2. The ALT company

As with the majority of all public elementary schools in Japan, the English classes at the elementary school were co-taught by an assistant language teacher (ALT) and the

Japanese homeroom teacher (see Martin, 2010 for an overview of the ALT industry). The ALT company, which had also provided previous ALTs to the same elementary school since 2006, viewed the oral communication proficiency test as a potential assessment of its ALT staff. It also viewed the oral communication proficiency test development project as a way to promote the company's educational services and senior staff, who worked together on this project as the development team. The senior staff consisted of three experienced, native English speaking teachers, one of whom was designated as the project leader. These three teachers worked together to develop, administer (i.e., interview) and score (i.e., rate) the participating students.

3. 1. 3. The project advisor

Lastly, as advisor to this project, it was my duty to train the project team with regard to oral communication testing. I was involved in the entire project from its creation to implementation and subsequent evaluation.

3. 2. Test purpose: Achievement versus proficiency

At the beginning of the project, it was unclear if the principal wanted a test designed to assess the students' language achievement or language proficiency. The difference between the two types of tests was also a point of confusion among the project team. A criterion-reference test (CRT) focuses on how much of a specified language skill or area of language knowledge a student has learned whereas a norm-reference test (NRT) aims to compare students' outcomes to each other (Brown, 1988). For instance, in order to check to see how many new vocabulary words a student has learned, a teacher would administer a CRT, i.e., an achievement test. One would hope that all students do individually well, and such tests are usually designed with the hope of 100% achievement outcomes for all test takers. On the other hand, in order to place a student in an appropriate class level or assess her overall language ability, she should take an NRT, i.e., a test of proficiency. The outcome of a proficiency test shows an individual's position in relation to other students with regard to a defined standard of ability. Another important difference between the two types of tests is that though achievement tests (i.e., CRTs) are used throughout a course in order to provide a final course grade, they cannot be equated to define a student's language proficiency. However, proficiency tests (i.e., NRTs) can be used to show a student's change in proficiency over time.

Upon understanding the difference between achievement and proficiency tests, the principal decided that a proficiency test to assess the students' language use would be the best test to administer. He said that he wanted to develop language standards for his school, and so he asked the project team to develop language standards and conduct a trial test based upon those standards.

Thus, this project aimed to define what an oral communication proficiency standard could be for Japanese upper elementary public school children, to create a performance-based test upon those standards, and to implement a pilot version of the test. In short, this study aimed to address the needs expressed by a school principal while at the same time it also aimed to fill a gap in the field of oral communication proficiency testing of young EFL learners.

4. Research Questions

Therefore, in order to investigate upper elementary school students' oral communication proficiency, this pilot study is an evaluation of an original oral communication proficiency standard and its implementation. The oral communication proficiency standard was operationalized in a rubric and used by three raters to assess elementary school students. Thus, the evaluation of this project was based upon the following research questions:

1. To what degree do the students fit the oral communication proficiency standard as defined by the created rubric?
2. To what degree do the categories of the rubric identify distinct areas of oral communication proficiency?
3. To what degree are the three raters consistent with each other in judging students' oral communication proficiency?

5. Methods

5. 1. Participants

This study involved 44 Japanese 4th-grade elementary school students. The students belonged to the same school and came from two homeroom classes (n = 23 and n = 21). Of the 44 students, 8 students were excluded from the rating process. Six student assessments were used for rater training, one was absent on the test day and one student was considered to be a balanced English-Japanese bilingual, i.e., not an EFL student. Thus, 36 Japanese 4th-grade elementary school students were assessed.

All students received two 45-minute English classes per week led by the same ALT who co-taught with each respective Japanese homeroom teacher. In 2006, the school provided English classes once a week for all grades 1 to 6 over the entire 35-week academic year, and in April 2007, the school increased the number of lessons to twice a week. Though the majority of the students had probably entered the school as 1st-grade students in 2007, it is unknown just how many of the students had received two English lessons a week between April 2007 and September 2010, for an approximate total of 340 class hours.

The three project team members were also considered to be participants of this study. They conducted interviews and rated all 36 students. None of the project team members had received training with regard to foreign language assessment prior to this project.

5. 2. Instruments

5. 2. 1. Rubric

Rubrics are used to define the quality of a student's performance across categories. Rubrics should identify the categories to be evaluated and define the optimal as well as the lowest levels of expected student performance (Curtain & Dahlberg, 2004). For this oral communication proficiency test, the project team focused on three categories: communicative competence, vocabulary/syntax, and interactional competence (see Table 1).

Communicative competence reflected the student's ability to use language in order to achieve the goal of the task, which in this case was to share information (see a full

Table 1. Rubric: EFL Oral Communication Proficiency for Upper Elementary School Students

	4	3	2	1
Communicative Competence	Able to communicate the expected amount of information.	Able to communicate most of the expected amount information.	Able to communicate some of the expected information.	Able to communicate minimal information.
Vocabulary/ Syntax	Displays a variety of syntax and vocabulary.	Displays some variety of syntax and/or vocabulary.	Displays a more limited variety of both syntax and vocabulary.	Displays little syntax or vocabulary.
Interactional Competence	Responds appropriately to all input and is able to initiate interaction with the interviewer.	Responds appropriately to most input. Initiates some interaction with the interviewer. Some support on the part of the interviewer is necessary.	Responds appropriately to some input. May initiate sometimes. Some effort and support on the part of the interviewer occurs.	Responds appropriately to some input but constant effort and support on the part of the interviewer are necessary.

description of the task below). Vocabulary/syntax focused on the type and variety of language used during the task. Interactional competence assessed the student's ability to react appropriately to input and initiate language.

Proficiency-based rubrics are created for a population of learners, and as stated, because a language-use standard for upper elementary Japanese EFL students did not exist, the project team created the rubric to assess students' oral communication proficiency. The creation of the rubric was based upon classroom observations, teaching experience with other Japanese students of the same grade level within the same school district and knowledge of the teaching materials and methods in use. Thus, it was believed that the rubric appropriately reflected the EFL proficiency of Japanese elementary school students, grades 4 to 6.

5. 2. 2. Oral communication performance–based task

The oral communication performance-based task was designed for one student and one rater. Both the student and the test rater brought a photograph of a special occasion that included the test participant and at least two friends or family members to the test. During the course of the test, the student and the rater talked about their photographs, first the student's and then the rater's. The students were told to tell the rater about their photograph, answering any questions the rater had about the photograph, and then to talk about the rater's photograph. It was expected that the most common aspects of such a photograph to talk about were (a) the people in the photograph, (b) the location depicted in the photograph and (c) the temporal aspect of the situation in the photograph or of the related special occasion. Therefore, the three aspects of people, location and time were used to assess shared information (communication competence), the English used to share the information (vocabulary and syntax), and the interaction with an interlocutor (interactional competence).

This task was believed to cover many of the aspects of a performance-based language-use task. It was authentic in that the photographs were real and related to the participants' lives. The raters' photographs had the appeal of being about a foreigner's life and a foreigner's immediate circle of friends or family. In addition, the task was designed to be interactive. Conversations about photographs spark statements, questions, responses of surprise as well as potential commonalities.

From a language point of view, it was believed that the students should have been able to provide basic information about people, location and time. It was also believed that students should have been able to initiate language, respond to and ask questions.

In sum, this task was believed to match the characteristics of the group of learners for which it was made in addition to their foreign language abilities. Lastly, this task was believed to have met the three basic requirements of performance assessment in that (a) students had to perform a task, (b) the task was authentic and interactive and (c) each student's performance was to be rated by more than one rater (Brown & Hudson, 1998).

5. 3. Procedures

This study was conducted in September 2010, approximately five months into the academic year. Because the school principal expected to assess a greater number of students in the future, each student assessment was limited to 5 minutes for practical purposes. Each homeroom class was scheduled to have two English classes in a row during the test day's class schedule. With this arrangement, one entire class of students was tested on the same day and two days later the other class of students was assessed. The tests were conducted in the school library, and for pilot test purposes, all tests were videotaped.

Approximately 2 weeks prior to the first test day, the school principal met with the project leader and the two Japanese homeroom teachers. The school principal outlined the goals of the test, and said that the students were not to be told that their English language use was to be assessed. Because this project sought to define a standard of oral communication proficiency for upper elementary school children, the mention of an assessment may have motivated students to study and/or induced test anxiety which may have distorted the test outcomes. The mention of an assessment may have also caused confusion or concern among the teaching staff and parents. Students were only told that they would have a special lesson during which three guest native English speakers would also come and that they would each get the chance to talk with one of the guest native English speakers about their prepared photograph.

On the test day, another member joined the project team. This member was Japanese and acted as an intermediary. She spoke only in Japanese to the students. Her job was to collect the three students called by the homeroom teacher, ensure that they had their photographs and walk with them to the school library. On the way, she was to encourage the students to not wait to be questioned, but to talk as much as they could in English and have fun. They were told that once they finished they were to go back to their classroom.

5. 4. Data analysis

This study employed the use of Rasch analysis (Rasch, 1966) as opposed to Classical Test Theory (CTT). CTT does not make allowance for the difference in difficulty between items, and most notably for this study, the difference of severity among raters (Bond & Fox, 2007). In traditional tests, one test item may be more difficult than another, and thus should carry greater weight rather than simply be counted as correct as an easier item. The same can be said of the difference among raters, whether with regard to a written essay or an oral interview. If one rater judges a participant more severely (or more leniently) than other raters, the participants' scores should be adjusted for this influence. Instead of ignoring this inherent issue, the Rasch model takes each facet into consideration and reports on each facet separately. Furthermore, in test development and rater training, it is important to ascertain each rater's tendencies.

Many-facets Rasch model (MFRM) allows for data diagnostics from multiple perspectives. In this pilot study there were three facets: persons (the students), items (the rubric categories), and the raters. All of these facets are of great importance with regard to the development of an oral communication proficiency test. MFRM analysis shows if the student population matched the oral communication proficiency model, if the rubric categories represented an appropriate difficulty range, and if raters, individually as well as a group, appropriately and consistently scored each student based upon the rubric (Bachman, Lynch & Mason, 1995).

After all the interviews had been conducted, the video footage of the 36 interviews was downloaded to a computer and burned on to DVDs. The raters independently scored all 36 interviews based upon the rubric (Table 1), scoring each student on each category on a scale from 1 (low) to 4 (high). The raters submitted their scores via a Microsoft Excel file prepared by the project advisor. The project advisor then combined all of the data and imported it into Minifac 3.67.0 (Lincare, 2010), a free, but limited, version of the Rasch statistical software Facets.

6. Results

In inferential statistics, the commonly reported reliability estimate is Cronbach's alpha whereas in MFRM analysis the analogous statistics is called the separation reliability (Wright & Masters, 1982). Separation reliability is based upon variance in the data and is represented on a scale between 0 and 1.0, with 0 reliability indicating complete randomness and 1.0 indicating perfect reliability.

In addition, MFRM analysis also provides a separation index, which is based upon the standard deviation of the data (Wright & Masters, 1982). The separation index shows to what degree the data is spread out. In the case of persons, a high degree of separation would indicate that some people scored low while other scored high. In the case of items, a high degree of separation would indicate that the items covered a wide range of difficulty, i.e., some items were easy while others were difficult. In the case of raters, a high degree of separation would indicate a range of rater severity, i.e., some raters were lenient while others were severe. A high degree of separation would be ideal for persons and items, but not raters (Bond & Fox, 2007). One would hope that raters would judge with the same degree of severity, and therefore, have a low degree of separation.

6. 1. Persons

The person separation reliability (.91) indicated that the test was highly reliable in differentiating among person ability and that if tested again at the same difficulty levels, the students would probably score in similar fashion. The person separation index (3.17) underscored that the students were spread out to a great extent along the scale of item difficulty. The significant *p* value ($p < .00$) indicated the rejection of the null hypothesis that all students had the same ability.

While the person reliability estimates indicated that the students had different degrees of ability, further investigation into the students' scores focused on student variance. Investigation into student variance showed to what degree students fit the model of the study.

Mean square, fit statistics are based upon the variation within the data. Infit Mean Squares represent the observed data while the Outfit Mean Squares represent each data point among the pattern of all data; mean square statistics range from 0 to positive infinity with an expected value of 1.0 and are considered acceptable when between 0.75 and 1.3 (Bond & Fox, 2007). Infit and outfit *z*-scores have an expected value of 0 and are acceptable when between +2.0 and -2.0. (Bond & Fox, 2007).

Table 2 shows the fit statistics for all 36 students. The table is organized by the Infit Mean Square and the students are represented by a student number. Only 15 of the 36 students fell within the acceptable infit and outfit ranges. Ten or 28% of the students had fit statistics greater than 1.3 which indicated underfit, i.e., there was too much variation in the judgment of their scores. Eleven or 31% of the students overfit the model, which means there was too little variation among their judged scores.

Three of the mis-fitting students, 11, 25 and 27, also had *z*-scores beyond the acceptable range. Table 3 shows the raters' scores for each of these students. Rater 1 judged Student 11 much more severely than Rater 2 and Rater 3. However, Rater 2 judged Student 11 severely on communication competence. It was the disparity among the raters, not the ability of the student, which defined Student 11 as mis-fitting the model. Students 25 and 27 received the exact same ratings by all raters. Given the amount of disparity among the raters it would be improbable that the raters would converge to such an extent as they did for Students 25 and 27, which led to the unacceptable *z*-scores.

Table 2. Person fit statistics

Student	Infit Mean Square	Outfit Mean Square	Infit *z*-score	Outfit *z*-score
11	3.35	3.56	2.41	2.44
9	1.52	1.45	0.95	0.82
2	1.44	1.53	1.06	1.10
22	1.38	1.48	0.89	1.01
19	1.37	1.48	1.01	1.16
5	1.35	1.32	1.10	0.93
10	1.32	1.34	0.67	0.68
14	1.32	1.36	0.68	0.69
15	1.32	1.34	0.67	0.68
1	1.27	1.10	0.68	0.38
35	1.25	2.06	0.55	1.07
28	1.04	0.97	0.23	0.10
8	1.00	1.00	0.00	0.00
18	1.00	1.00	0.00	0.00
20	1.00	1.00	0.00	0.00
21	1.00	1.00	0.00	0.00
23	1.00	1.00	0.00	0.00
26	1.00	1.00	0.00	0.00
34	1.00	1.00	0.00	0.00
6	0.97	1.05	0.07	0.27
24	0.92	0.89	−0.10	−0.10
29	0.92	0.89	−0.10	−0.10
13	0.90	0.87	−0.15	−0.20
3	0.85	0.75	−0.33	−0.44
30	0.82	0.84	−0.39	−0.26
17	0.79	0.69	−0.34	−0.43
31	0.71	0.54	−0.53	−0.56
36	0.71	0.54	−0.53	−0.56
12	0.64	0.31	−0.27	−0.40
16	0.58	0.52	−1.24	−1.18
33	0.58	0.47	−0.60	−0.74
7	0.53	0.43	−0.98	−1.03
32	0.51	0.48	−1.83	−1.71
4	0.30	0.21	−1.41	−1.51
25	0.11	0.09	−2.06	−2.04
27	0.11	0.09	−2.06	−2.04

Table 3. Rater's scores for students 11, 25, 27

Rater	Student 11			Student 25			Student 27		
	C.C.	V.S.	I.C.	C.C.	V.S.	I.C.	C.C.	V.S.	I.C.
1	2	1	2	2	2	2	2	2	2
2	1	3	2	2	2	2	2	2	2
3	2	3	2	2	2	2	2	2	2

Note: C.C = communicative competence; V.S. = vocabulary/syntax; I.C. = interactional competence

6. 2. Items

The item separation reliability (.73) indicated that communication competence, vocabulary/syntax, and interactional competence were only slightly different regarding task difficulty, i.e., none of the categories were much more or much less difficult than the others. The item separation index of 1.65 verified that the items were somewhat separated along the difficulty continuum, but not to a large degree. The significant *p* value (*p* < .00) indicated the rejection of the null hypothesis that all items had the same difficulty.

Table 4 shows the fit statistics by item, i.e., the rubric categories. Fit statistics for both communication competence and vocabulary/syntax were within the acceptable ranges. Outfit Mean Squares for interactional competence, on the other hand, was below the acceptable range of 0.75, which indicated there was too little variation among the students' scores, i.e., the raters' judgments.

Table 4. Item fit statistics

Item	Infit Mean Square	Outfit Mean Square	Infit *z*-score	Outfit *z*-score
Communication competence	1.11	1.21	0.78	1.06
Vocabulary / Syntax	1.11	1.12	0.72	0.62
Interactional competence	0.75	0.64	-1.64	-1.88

6. 3. Raters

As mentioned, a high rater separation reliability for raters does not mean a high degree of consistency among raters. Rather, a high reliability would indicate that raters were consistently different with regard to the severity of their ratings (Bachman *et al*., 1995). The rater separation reliability was .88 and the rater separation index was 2.77, indicating a moderate amount of consistent disagreement among the raters. The significant *p* value (*p* < .00) indicated the rejection of the null hypothesis that all raters scored the students at the same level of severity.

Fit statistics regarding raters represent to what extent each rater was self-consistent in scoring all students. Table 5 shows that all of the fit statistics fell within the acceptable ranges, indicating individual rater consistency.

Table 5. Rater fit statistics

Rater	Infit	Outfit	Infit	Outfit
	Mean Square	Mean Square	*z*-score	*z*-score
1	1.03	0.98	0.26	-0.01
2	1.02	1	0.17	0.06
3	0.94	0.99	-0.36	-0.01

6. 4. Bias analysis

In CTT, significant statistical results are often checked for interaction among variables. MFRM analysis can also highlight effects of interaction between facets through MFRM bias analysis. Bias analysis detects any deviation from the expected patterns between facet relationships.

Bias analysis between person by item, i.e., student by category, showed that there was evidence of an interaction for at least one category of the rubric for 32 of the 36 students. This means that most students' scores, i.e., their judged ratings, varied unexpectedly from the expected outcome. Figure 1 shows the degree of severity by category for each rater. All raters scored communication competence differently, and there is a clear difference between Rater 1 and Rater 2. All raters converged, yet did not agree with regard to vocabulary and syntax. Yet although Rater 2 and Rater 3 judged interactional competence in seemingly identically fashion, Rater 1 was much less severe.

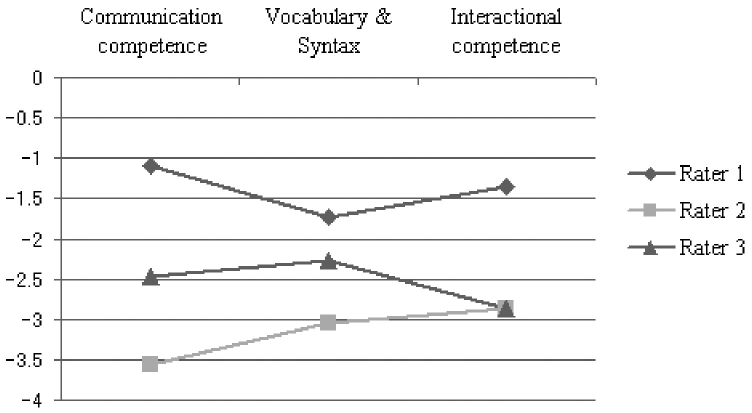

Figure 1. Rater by item bias

7. Discussion

This study set out to create an EFL oral communication proficiency standard and to then use the standard to assess 4th-grade elementary school Japanese students. Rasch analysis was used to answer this study's three research questions.

7. 1. Student fit

The interview test reliably dispersed the students across a range of difficulty, but only less than half of the students fit the oral communication proficiency standard model. Invariably, this was somewhat due to the disparity among the raters' judgments. However, there are other factors that may have caused student misfit with this model. First, as mentioned, it was decided that the students would not be told that their language ability was going to be tested. Perhaps if they had known about the evaluation in advance, they may have been more ready to actively participate and use their English language ability to a higher potential. Second, the students had never experienced such a one-on-one conversation before. Not only did they not expect to be assessed, but they also had no prior experience in such a situation, and thus, did not know what was expected of them. Third, culturally speaking, 10 year-old Japanese children may not have had the experience of speaking with an adult stranger, regardless of nationality and language, about such a personal topic as a special occasion with friends or family. In short, a general lack of experience upon the part of the student may have had a considerable influence on each student's ability to perform.

7. 2. The rubric

The item analysis showed that the rubric categories were relatively similar in difficulty. During the creation of the rubric, the concept of category difficulty did not arise. That is, the categories of communication competence, vocabulary/syntax and interactional competence were not created with the belief that one was more difficult than the other. They were created in order to find out to what extent students would be able to show their ability to convey information, to use a variety of vocabulary and syntactical forms and to what degree they could interact with an interlocutor. On the one hand, the difference in rater severity cannot be ignored. However, a review of the video footage also showed that the raters did not attempt to develop the students' use of these three categories.

The task design focused on a photograph depicting a special occasion and talking about the associated people, location and time. The interactiveness of this task did not constrain the participant and also provided opportunities for support (Liddicoat, 1997).

Yet instead of finding out if students could elaborate on any of the topics, the raters treated the task more like an information gap activity. For example, upon hearing that the photograph was taken in a restaurant, the rater then moved the conversation toward the people or time of the photograph rather than asking more about the restaurant, e.g., "Is it your favorite restaurant?", "Do you go there often?", "What kind of restaurant is it?" Moreover, each rater gave a fairly complete explanation of their own photograph, which did not extend much opportunity to the students to engage the rater in conversation.

7. 3. Rater consistency

The raters were consistent in how they individually scored all 36 students. However, inter-rater consistency was not evident. Of all aspects of this pilot project, rater consistency and understanding how to apply the rubric were the most important. The results of the pilot project will be used for subsequent rater training and development of the raters' interviewing skills.

The bias analysis between person and item clearly showed the influence of the disparity among the raters. If such a result occurred on a paper-and-pencil test it would indicate that a student either performed better than expected or worse than expected. Yet in this case, students were assessed by raters. Rather than performing better or worse than expected, the raters judged student performance either better or worse than expected. This result is a reminder that the raters significantly assessed the students in varying degrees of severity.

However, another way of viewing the data other than from the perspective of raters' severity is student lack of ability. Though the raters did not show inter-rater reliability, they did have severity in common. As stated, this study was designed for upper elementary school students and 4th-grade students are at the lower end of this range. Perhaps evaluating 5th and 6th-grade students using the same task and language standards, raters would show less rater severity overall and a more appropriate understanding of students' ability levels.

8. Limitations and suggestions for future studies

This study is one of the first of its kind to take place in Japan with public elementary school students. Though this can be considered a strength, it is also a clear limitation. Without previous research to learn from or replicate, pitfalls cannot be avoided. This study has at least three key limitations.

First, this study attempted to create an EFL oral communication proficiency standard. Any such standard must be continually re-evaluated and re-assessed over a number of trials before it can be used with confidence. In dealing with young EFL learners, the standard can only be a skeleton version of what language competence is known to be. As such, the definitions between the three categories may not yet be precise enough to use in assessment. Second, the project team was not experienced in EFL language assessment. This lack of experience coupled with a new instrument more than likely contributed to the lack of inter-rater reliability. Further rater training is necessary. Lastly, due to time constraints, this study used a 5-minute oral communication performance-based task. Five minutes is much too short to adequately assess a student's language-use proficiency. This time constraint also limited the variety of language that could be assessed.

Future studies need to build upon this effort to define what language proficiency means for the population of young EFL learners. Future studies also need to focus on the training of raters. I would also encourage future researchers to secure more time and to develop alternative task types. It is also important that future studies do not rely upon CTT. Rasch analysis offers a much greater and broader understanding of data that is based upon human judgment.

9. Conclusion

The language assessment of young EFL learners is coming. Over the last decade, EFL classes in Japan at the elementary school level have steadily increased in hours and coverage of the nation's public schools. As compulsory language education is set to begin, more and more focus will be on the outcomes of elementary school EFL classes. If MEXT continues its decentralized EFL policy, it will be up to the boards of education, individual school staff members and outside agencies to create EFL standards.

However, as this study has shown, the development and use of such standards are not easy. Yet despite the limitation of this study, it has provided important insight to the field of children's EFL assessment. First, students are able to converse on topics in one-on-one situations. The results of this study also inform EFL educators that more exposure and expectation of language use during regular class hours is necessary. Second, the rating of students' language use abilities is possible. Students are able to do more than merely identify vocabulary words or only understand to language input. The students' ability to participate in a conversation allows for the ability to assess their participation. Experienced raters should be able to do so adequately and fairly given an appropriate rubric to follow. Lastly, Rasch analysis provides a clear understanding of how raters assess students, how students are dispersed along a range of difficulty and how facets interact.

Acknowledgements

I would like to acknowledge and thank the staff at the Institute for English Communication especially Mieko Itoh, Michael Carr, Stuart Neilson, Wendy Gedye and Keiko Taguchi. Without their hard work and assistance, this study would not have been possible.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675.

Butler, Y. G. (2004). What level of English proficiency do elementary school teachers need to attain to teach EFL? Case studies from Korea, Taiwan, and Japan. *TESOL Quarterly, 38*(2), 245-278.

Butler, Y. G. (2007). Foreign language education at elementary schools in Japan: Searching for solutions amidst growing diversification. *Current issues in language planning, 8*(2), 129-147.

Butler, Y. G., & Takeuchi, A. (2006). Evaluation of English activities at Japanese elementary schools: An examination based on junior step bronze test. *JACTEC Journal, 25*, 1-15.

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge: Cambridge University Press.

Cameron, L. (2003). Challenges for ELT from the expansion in teaching children. *ELT Journal, 57*(2), 105-112.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Council for Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Curtain, H., & Dahlberg, C. (2004). *Languages and children: Making the match*. Boston: Pearson Education.

Dunlea, J., & Matsudaira, T. (2009). Investigating the relationship between the Eiken test and the CEFR. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR: Research perspectives* (pp. 103-109). Arnhem: European Association for Language Testing and Assessment.

Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing, 17*(2), 261-277.

Johnstone, R. (2000). Context-sensitive assessment of modern languages in primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing, 17*(2), 123-143.

Katsuyama, H., Nishigaki, C., & Wang, J. (2008). The effectiveness of English teaching in Japanese elementary schools. *RELC Journal, 39*(3), 359-380.

Linacre, J. M. (2010). Minifac (version 3.67.0) [computer software]. Chicago: MESA Press.

Long, M. H., & Porter, P. A. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly, 19*(2), 207-228.

Martin, R. (2010). Team-teaching in Japanese public schools: Fissures in the alt industry. *Language, culture, and communication: Journal of the College of Intercultural Communication, 2*, 145-152.

McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.

Ministry of Education, Culture, Sports, Science and Technology. (2001). 小学校英語活動実践の手引き *[Practical handbook for elementary school English activities]*.Tokyo: Kairyudo Publishing.

Ministry of Education, Culture, Sports, Science and Technology. (2002). Developing a strategic plan to cultivate Japanese with English abilities. Retrieved December, 30, 2003, from http://www.mext.go.jp/english/news/2002/07/020901.htm

Ministry of Education, Culture, Sports, Science and Technology. (2003). The course of study for foreign languages. Retrieved March, 14, 2006, from http://www.mext.go.jp/english/shotou/030301.htm

Ministry of Education, Culture, Sports, Science and Technology. (2008, June). 小学校学習指導要領解説: 総合的な学習の時間編 [Elementary school course of study explanation: Comprehensive studies].

Munoz, C. (Ed.). (2006). *Age and the rate of foreign language learning*. Clevedon: Multilingual Matters Ltd.

Nikolov, M. (2009). The age factor in context. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 1-37). Berlin: Walter de Gruyter.

Nikolov, M., & Curtain, H. A. (2000). *An early start: Young learners and modern languages in Europe and beyond*. Strasbourg: Council of Europe.

Nikolov, M., & Djigunovic, J. M. (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual Review of Applied Linguistics, 26*, 234-260.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawai'i Press.

Nunan, D. (1989). *Designing tasks for the communicative language classroom*. Cambridge: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago.

Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing,* pp. 115-122.

The Society for testing English proficiency. (n.d.). Retrieved October 20, 2010, from http://

stepeiken.org/

The Society for testing English proficiency: Jr. Step Bronze. (n.d.). Retrieved October 20, 2010, from http://stepeiken.org/

The Society for testing English proficiency: Timeline. (n.d.). Retrieved October 20, 2010, from http://stepeiken.org/

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.