Developing a reading test based on CEFR-J

Yasuko Ito

Abstract

The purpose of this study is to discuss the development of a reading test based on the CEFR-J. The research question is whether or not we can observe (1) rank-ordering of test items or tasks with calibrated item difficulty from the lower A1.3 to the intermediate B1.2 levels, and (2) item clusters under the same level specifications in terms of their difficulty and the distinctiveness of the cluster from those of other levels'. A reading test was developed following the descriptors of the CEFR-J and given to freshmen at a university in Japan. The test validation will be discussed based on the results obtained from the test administration.

Introduction

Different kinds of standardized English tests have been used in secondary as well as tertiary education in Japan. There is also a discussion of their use as university entrance examinations. Among some advantages, one is that their scores can be interpreted in a wider context, by comparing them with other test takers', while in-house tests do not allow that. On the other hand, a disadvantage could be that some tests can be too challenging for a certain group of students, and the tests may not measure their proficiency accurately or discriminate them well enough. One possible solution would be to develop a test that is suitable for the students, although it is a time-consuming task. Using level specification as a basis for the test development may be potential. One such specification is "Common European Framework of Reference for Languages: Learning,

teaching, assessment" (Council of Europe, 2001), also known as CEFR.

There is a widespread of the use of the CEFR across the world. It has been widely utilized in many educational contexts. Japan is no exception. However, it has been claimed that the original CEFR is not suitable for the Japanese contexts in that the level allocation does not appropriately reflect the proficiency of Japanese learners, the majority of whom are likely to fall into lower levels of the CEFR. They are placed somewhere in A, and CEFR, therefore, does not discriminate learners properly.

The CEFR-J (Tono, 2013), CEFR which was adapted to the Japanese context, was developed and it divides A and B levels into more levels. Since it became available, it has been referred to in various research as well as pedagogical contexts. It has been applied in different settings, and one of the trials is to develop a test based on the CEFR-J.

The purpose of this paper is to discuss the development of a reading test based on the CEFR-J. A reading test was developed following the descriptors of the CEFR-J and given to freshmen at a university in Japan. The test validation will be discussed based on the results obtained from the test administration.

CEFR and CEFR-J

According to Council of Europe (2018), the CEFR "was designed to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabuses and curriculum guidelines, the design of teaching and learning materials, and the assessment of foreign language proficiency." One of the well-known aspects of the CEFR is can-do statements. Can-do statements specify what a learner at a certain level "can do" with the language. There are six levels in the CEFR: A1, A2, B1, B2, C1, and C2. Those at the A1 and A2 levels are considered to be basic users, while B1 and B2 learners are independent users, and C1 and C2 learners are proficient users. The CEFR provides descriptors of four language skills for each level.

One problem with using the CEFR in a Japanese context is that 80% of the Japanese learners of English belong to the A level on the CEFR (Tono, 2013). Therefore, it does not discriminate the Japanese learners well enough. The CEFR-J, the CEFR adapted to the Japanese context, tries to solve this problem by dividing the A and B levels into smaller categories. Table 1 shows the levels given in the CEFR and the CEFR-J.

Table 1

Comparison of Levels in the CEFR and the CEFR-J

CEFR		A1	A2	B1	B2	C1	C2
CEFR-J	Pre-A1	A1.1	A2.1	B1.1	B2.1	C1	C2
		A1.2	A2.2	B1.2	B2.2		
		A1.3					

Tono (2013) claims that Pre-A1 is equivalent to the level of foreign language activities at elementary schools, A1 to the introductory level of junior high schools, A2 to the level of second and the third years of junior high schools, B1 to that of the completion of high school English and the tertiary level, B2 to the level of the liberal arts at a university, and C to the level of native speakers or higher.

For each level in four different skills, descriptors in a can-do statement format are provided. For example, descriptors for the A2.1 reading is, “簡単な語を用いて書かれた人物描写、場所の説明、日常生活や文化の紹介などの、説明文を理解することができる” (One can comprehend an explanatory passage, such as a description of a person or a place, and an introduction of a lifestyle and a culture, written with simple vocabulary) (Tono, 2013). Such a descriptor would allow us to design and implement language classes as well as assess students’ language proficiency. However, we should also be aware of the challenges posed by different researchers. In the next section, such

challenges will be reviewed.

Concerns Regarding the CEFR and the CEFR-J

The use of assessment scales such as the CEFR and the CEFR-J can provide language educators with firm bases for program and test development. Yet, a number of researchers have voiced concerns about them. One concern was raised by Alderson, Figueras, Kuijper, Nold, Takala, and Tardieu (2006), who examined the usefulness of the CEFR for the reading and listening test construction. They found some problems with it, such as the lack of definition with many terms. Another concern is about the illustrative nature of the CEFR. Hulstijn (2007) claims "that the CEFR is not based on empirical evidence taken from L2 learner data" (p. 666).

Although still limited in amount, some investigation has been done on the CEFR-J. Runnels' study (2014) explored the reliability and content of the CEFR-J's A-level can-do statements. She found some problems with the A-level descriptors in terms of the reliability, and she attributed the finding to two factors. One was the content of the descriptors, and the other was the learner characteristics.

The purpose of the current study is, therefore, to address such concerns by empirically examining the hierarchy of reading ability depicted in the CEFR-J reading scales.

Research Questions

The purposes of this study are to examine whether or not the following can be observed: (1) rank-ordering of test items or tasks with calibrated item difficulty from the lower A1.3 to the intermediate B1.2 levels, and (2) item clusters under the same level specifications in terms of their difficulty and the distinctiveness of the cluster from those of other levels'.

As a first step of the study, a test was developed that is targeted to the students of Kanda University of International Studies, located near Tokyo, Japan, where the author teaches. Developing and administering the test to the actual students enables us to examine to what extent the implementation of the CEFR-J in test development is feasible and what challenges we may face.

Method

The entire process of the study consisted of four stages: Test development, pilot testing, test administration, and the evaluation of the developed test. Each stage will be described in detail.

Development of a reading test

The first stage of the study was to develop a reading test based on the CEFR-J descriptors. Rather than creating test items for all levels from Pre-A1 through C2, five levels were selected as target levels in the current study among 12 levels given in the CEFR-J, namely, A1.3, A2.1, A2.2, B1.1, and B1.2. These five levels were identified as those the majority of the students at the target university would fall into, based on the author's own teaching experience.

Test development had five phases: Familiarization, material collection, passage evaluation, passage selection, and test construction. Five researchers, including the author, took part in this process of the test development. In the first phase, Familiarization, we familiarized ourselves with the descriptors of each of the five levels, and we often found the interpretation of the descriptors challenging. For instance, the descriptor of A1.3 includes expressions such as "topics that one is personally interested in, such as sports, music, travel," "easy vocabulary," and "passage." It was not easy to determine what kind of vocabulary is considered to be "easy," for example.

Material Collection was the next phase in which we collected reading passages according to the descriptors of each level. For example, for A2.2, whose descriptors state "practical and concrete materials whose contents are easily predicted, such as travel guidebook and recipe," online travel guide and recipe were collected. Likewise for B1.1, which says "simple instructions, such as how to play games, how to fill out a registration form, and how to assemble a material," instructions of a game were chosen. Sources of collected materials include short stories, magazines, online news articles, and educational institution websites, and from these sources, 27 passages were collected in total.

In the third phase, Passage Evaluation, we evaluated the collected passages. In order to ensure the interrater reliability, four of the researchers evaluated 27 passages, using a Manual for Reading Passage Judgment, which was prepared by one of the researchers. The judgment was based on such questions as, "does this passage surely belong to the intended level?" or "does the text really represent the text type/genre mentioned in the CEFR-J?" All passages were evaluated by the researchers on a 4-point scale: Strongly agree (1), Agree (2), Disagree (3), and Strongly disagree (4).

Based on the judgment ratings, passages whose levels were agreed by all members were selected, which was the fourth phase, Passage Selection. Out of the 27 passages, 12 passages were eventually selected as those to be used in the test. Here is an example of a reading passage for A2.1, whose descriptor says "One can comprehend an explanatory passage, such as a description of a person or a place, and an introduction of a lifestyle and a culture, written with simple vocabulary."

Universal Studios Japan

Universal Studios Japan (USJ) was the first theme park under the Universal Studios brand to be built in Asia. Opened in March 2001 in the

Osaka Bay Area, the theme park occupies an area of 39 hectares and is the most visited amusement park in Japan after Tokyo Disney Resort. Universal Studios Japan currently has eight sections: Hollywood, New York, San Francisco, Jurassic Park, Waterworld, Amity Village, Universal Wonderland and The Wizarding World of Harry Potter. Visitors are able to enjoy many amusement rides, ranging from child-friendly carousels to thrilling roller coasters and simulators, based on popular movies such as Spiderman, Back to the Future, Terminator 2 and Jurassic Park. The theme park also offers many opportunities to take pictures with popular characters' mascots such as Snoopy, Hello Kitty and the puppets of Sesame Street. There are also various shows put on every day, including a night parade whereby illuminated floats are paraded through the streets. (http://www.japan-guide.com/e/e4021.html)

In the final phase, Test Construction, we developed test items that reflect task features at each level. The questions were all multiple-choice questions, with four options. The following is an example of the question for the example passage above.

Question: According to the passage, which of the following is true of Universal Studios Japan?

1) It is the most visited amusement park in Japan.
2) It has nine sections.
3) It has many rides built based on popular movies.
4) It offers limited opportunities to take pictures with popular characters' mascots.

The correct answer is (3). Two to three items were created for each of the 12 passages, and 32 items were prepared in total for the 12 passages.

Pilot testing and test revision

In order to examine whether the prepared test can work properly, a pilot test was administered. Seventy-five freshmen from two different universities in Japan participated in the pilot study. Two forms, Forms A and B, were created with six passages and 16 items each. Among the 75 students, 39 took both Forms A and B for common person equation. Classical and IRT test/item statistical tests were run. This resulted in deleting two items and revising some items for distractor efficiency. The final version of the test had 11 passages, with 30 items.

Participants

The revised test was administered to 412 freshmen, consisting of 126 male and 286 female students. The students took a TOEFL ITP test as part of the school requirement, and their score ranged from 370 to 550.

Data analysis

The data were analyzed using classical test statistics as well as IRT item statistics using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). The 3-parameter was employed for the analysis as it fitted the data best among other models.

The data were analyzed first to examine the statistical quality of the test, and then, item parameters were calibrated using the 3-parameter model. The resulted parameter values were examined to address the research questions, i.e., if the test items are rank-ordered as predicted with the calibrated item difficulty from A1.3 to B1.2 levels, and if the texts and their items of the same sub-levels are observed under the same difficulty

cluster. When the texts and/or the items exhibited undesirable rank-order patterns, an attempt was made to elucidate what aspects of the scale descriptors, if any, were responsible for such patterns, for example, lexical difficulty, terminology problems, descriptive gaps, undistinguishable operations, and inconsistencies within and across different levels or lack of definitions.

Results and Discussion

In order to check the statistical properties of the data, descriptive statistics are calculated and reported in Table 2. The distribution of the data appears relatively normal as indicated by the centrality as well as dispersion indices such as Kurtosis (= -0.07) and Skewness (= -0.61). The reliability coefficient (i.e., Cronbach's alpha), however, falls at the relatively lower limit (= 0.67) of the acceptable range. This low reliability may be due in part to the large number (k=11) of varying texts included in the test, as it makes difficult to keep the dimensionality of the reading construct consistent.

Table 2

Descriptive Statistics

Mean	Median	*SD*	Kurt.	Skew.	Range	α
20.34	21	3.93	-0.07	-0.61	7-28	0.67

In order to answer the research questions, the items were first grouped under their intended levels and then rank-ordered based on the values of their difficulty parameters. Figure 1 presents the result of this procedure. The straight line that goes through the difficulty logit values of individual items is a trendline that shows the incremental direction of the values. As the trendline confirms, a general progression of item difficulty is observable from the items at the lower to the upper levels.

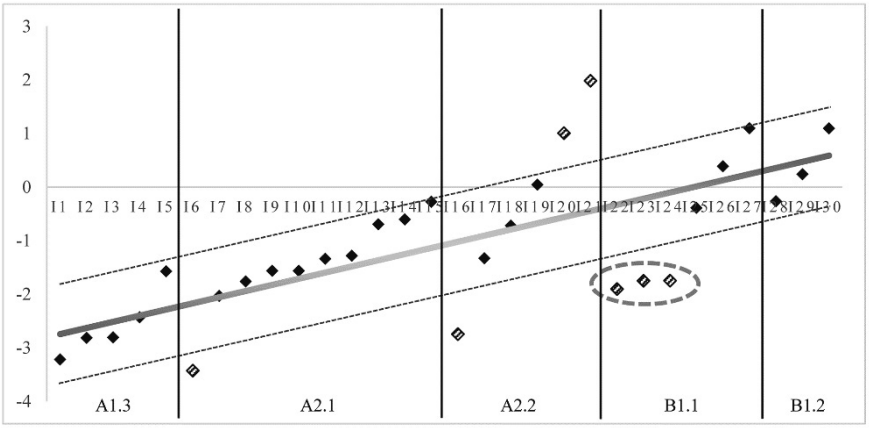

Figure 1. Item rank order

While the general progression is achieved with the individual items within and across sub-levels, there are still a few items that fall beyond the range of roughly ± one logit unit around the trendline. A closer look was given at these items to identify the possible causes of the marked deviance from their predicted difficulty. It appears that the cause of the deviant behaviors by these items is mostly due to the text type and the reading functions that the accompanying questions were to elicit.

For example, the three items in the dotted oval in Figure 1 were all based on the same reading passage for B1.1, the text type of which is recipe. Since the passage was a recipe, all the accompanying items were asking specific information through scanning (e.g., How many *Okonomiyaki* pancakes can you make using this recipe?). That is, the text type of a recipe imposed a limitation on the type of reading functions that the accompanying questions are to elicit.

Reading is an interactive process of text and function, and hence, the difficulty of a reading task is contingent upon how the two reading attributes interact with each other.

The recipe included in the test may, in fact, have been difficult to comprehend; however, as the reading functions that the items required to apply were not cognitively demanding, the overall difficulty level of the items also resulted in low. Consequently, the order of the average difficulty by individual sub-levels was affected by the difficulty of the test items that accompanied the recipe passage, making B1.1 easier than A2.2, as shown in Figure 2.

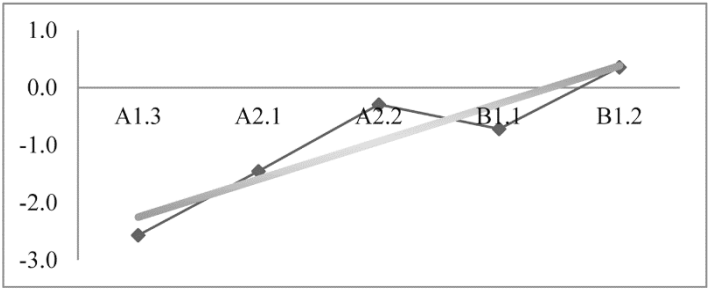

Figure 2. Difficulty progression by level

In order to confirm if the differences between adjacent levels were statistically meaningful, the mean difficulties of the five sub-levels were compared against each other using an ANOVA. The ANOVA analyses revealed the significant main effect ($F(3.27,1347.83)=291.3$, $p<.01$, $\eta^2=.35$), and the follow-up pairwise comparisons confirmed that the adjacent levels were all statistically different from each other. This result of the pairwise comparisons also meant that B1.1 was significantly easier than A2.2, the reversed order of the sub-levels from predicted.

It would be desirable if the source for the reversed order of the sub-levels could be identified. Using different types of reading passages may help reveal the cause. Once potential sources are identified, the rank-ordering will be in a predicted manner. Hulstijn

(2007) asserts, “there is no empirical evidence that all L2 learners at a given level beyond level A1 are able to perform all the tasks associated with lower levels” (p. 666). A better rank-ordering might, therefore, allow us to provide such evidence in the future.

As another future avenue of this study, the range of the examined CEFR-J levels can be expanded. The current study explored only five levels from A1.3 to B1.2 on the scale, but lower and higher levels may be included with students of a wider range of proficiency. Furthermore, one may examine the listening scale of the CEFR-J in a similar research manner, as L2 listening has not seen research efforts compared to L2 reading development.

Conclusion

This study examined if the use of the CEFR-J reading scales helps develop a level-specific reading test that includes the texts and their items pertinent to the level descriptors of the scales. The statistical findings informed that it may be feasible to develop such a level-specific reading test using the CEFR-J reading scales when at least two conditions are strictly met: 1) the entire process of test development is carefully coordinated, and 2) a close attention is given not only to the difficulty of the text types but also to the reading functions that can be realized using the texts and the level of their cognitive demands.

There is still much potentiality in this line of CEFR or CEFR-J research. A further investigation would shed light on the practicality and the application of the guidelines.

Acknowledgments

I would like to thank Prof. Tomoko Fujimura, Prof. Megumi Sugita, and Ms. Mariko Nomura, for their contributions to the test development, and Prof. Siwon Park for his support during the entire process of this study and the draft preparation.

This research is supported by Grant-in-Aid for Scientific Research (C) 16K02976 from Japan Society for the Promotion of Science.

References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly*, *3*, 3-30

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. Retrieved from the Council of Europe website: https://rm.coe.int/1680459f97

Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, teaching, assessment. Retrieved from the Council of Europe website: https://www.coe.int/en/web/common-european-framework-reference-languages/?

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal*, *91(4)*, 663–667.

Runnels, J. (2014). An exploratory reliability and content analysis of the CEFR-Japan's A-Level can-do statements. *JALT Journal*, *36(1)*, 69-89.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

投野由紀夫（編）（2013）『CAN-DO リスト作成・活用　英語到達度指標 CEFR-J ガイドブック』大修館書店