神田外語大学大学院紀要　言語科学研究　第25号（2019）【研究論文】pp. 57-71

Issues in developing English reading and listening tests based on the CEFR-J scales

Siwon Park (Kanda University of International Studies)
Megumi Sugita (Kanda University of International Studies)
Kento Inoue (Sundai College of Business & Foreign Languages)

Abstract

While the use of CEFR-J (Tono & Negishi, 2012) can provide firm bases for program development and test design, researchers have been skeptical of its excessive uses as its scales only deal with the functional side of learner performance. In this study, we report the results of a large scale project that examined the use of the CEFR-J Reading and Listening scales for test development. Through a rigorous multistep processes of test development, a set of English reading and listening tests were developed based on the CEFR-J scales. At the test development stages, it was first examined if the level descriptors include sufficient details to help test writers develop level specific test. Next, the tests were administered to a large group of EFL learners, and their scores were analyzed using IRT item analyses and the Bayesian hypothesis testing to examine if the rank-ordering of the carefully constructed test items is pertinent to their intended levels and the developed tests are level-specific. The findings indicate that while the level descriptors often resort to relative and personal expressions requiring subjective judgments as to the difficulty of a specific level, the development of level-specific reading and listening tests may still be feasible when the development procedures are carefully coordinated.

1．Introduction

The use of foreign language (FL) proficiency scales (or guidelines), as Table 1 shows, has become popular as they can serve various educational purposes such as curriculum development, test design, and program evaluation. While

some of them were employed for their intended uses in a clearly defined educational context, others such as the ACTFL guidelines and Common European Framework of Reference for Languages (CEFR) have been adopted or localized (e.g., CEFR-J) for their use in other educational contexts.

Table 1. Foreign language proficiency scales and guidelines

The United States	Europe	Australia	Japan
• FSI scale • ILR scale • ACTFL Guidelines	• ALTE Framework • Common European Framework of Reference for Languages (CEFR)	• Australian Second Language Proficiency Rating Scale	• CEFR-J(apan)

These FL scales are commonly employed to guide program development, instruction, and assessment; FL teachers may wish to employ such scales or guidelines in their construction of tests or syllabuses concerning real-life tasks (North & Schneider, 1998). More specifically, these scales can serve in teaching an FL (Negishi, Takada, & Tono, 2012; Nagai & O'Dwyer, 2011; North, Ortega, & Sheehan, 2010; Tono, 2013) as:

- a common basis for the development of L2 programs or curricula and function as a common yardstick for the evaluation of the program or the curriculum
- a reference point of learner progress at the predefined stages of long-term development
- a benchmarked guideline for examinee performance on a standardized L2 exam; i.e., they provide an interpretive guideline for score meanings in terms of can-do lists (Nagai & O'Dwyer, 2011)
- a set of guidelines from which tests can be built to suit local testing needs, when adapted with more elaboration (Davidson & Fulcher, 2007)

The use of the FL proficiency scales, therefore, suggest a great potential in FL education because they can provide firm bases for the development of language program, curriculum, and assessment. At a more global level, these scales can serve as a common yardstick for program evaluation *within* a system or as objective, comparable metrics *between* systems.

Despite such advantages, researchers have expressed concerns regarding the use of the scales in FL education (e.g., Bachman & Savignon, 1986; Spolsky, 1986; Pienemann & Johnston, 1987; Hulstijn, 2007; Runnels, 2013), and the following specifically concern the CEFR scales:

1) There is no guarantee that the progressive level distinctions and the number of levels are accurate, valid, or balanced (Lantolf & Frawley, 1988).
2) There is no guarantee that the level specific descriptors are accurate, valid, or balanced (North & Schneider, 1998).
3) These scales are often context-specific, and hence, the general use of a specific scale in a different context must be warned against (Spolsky, 1986).

The CEFR scales have been criticized because their descriptors are illustrative rather than normative. They are also language and context neutral rather than specific, and comprehensive rather than complete (North, Martyniuk, & Panthier, 2010). Being illustrative and context-neutral, however, the scales become open and flexible and can be adapted and localized to suit the intended purposes within and across different language contexts better.

The CEFR-J scales are an example of such adaptation for the educational use outside of the European context in which the CEFR was originally developed and used. The CEFR was selected and localized to develop a consistent educational system for foreign language education in Japan. As Table 2 shows, in the localizing process, a new level (Pre-A1) was added in the CEFR-J, and both A and B levels were further divided to include sublevels. These sublevels

were to help distinguish learners at the lower proficiency levels in the secondary education in Japan. There was no change made to C1 and C2.

Table 2. Comparison of CEFR and CEFR-J

CEFR		CEFR-J		
		Pre-A1		
Basic user	A1	A1.1	A1.2	A1.3
	A2	A2.1	A2.2	
Independent user	B1	B1.1	B1.2	
	B2	B2.1	B2.2	
Proficient user	C1	C1		
	C2	C2		

2. The CEFR and the CEFR-J for FL assessment

With respects to the use of the CEFR scales for FL assessment, they are classified as user-oriented rather than constructor-oriented, making it difficult to use them as rating scales or for the development of standardized tests (Fulcher, 2010; Hulstijn, 2007; North, 1991; Weir, 2005). Such a limitation explains why it is rare to find a study that adopted or adapted the CEFR for test development while there are a number of bench-marking studies to align the scores of a standardized test onto its descriptors. As Davidson and Fulcher (2007) argue, the CEFR scales may be used as a springboard to task and test development but not as a set of normative guidelines that can provide a direct reference of linguistic and cognitive functions to be tested. In exploring the possible use of the CEFR scales for the development of L2 standardized tests, therefore, one needs to employ a rigorous process of ensuring a number of parameters such as the context of use and theoretical rigor, coverage, and explicitness in relation to the descriptors.

The issues concerning the CEFR addressed so far are equally pertinent to the CEFR-J especially with regards to its use for test development. The

CEFR-J has been proposed to serve as a reference for curriculum and teaching materials development and for assessment in Japan (Tono, 2013). The CEFR-J has gone through a number of validation processes already, and several key materials have been prepared such as can-do lists and level specific wordlists. Yet, no scientific attempt has been made to explore the possible use of the CEFR-J scales for test development. Given the earlier concerns for the characteristics of the CEFR, the CEFR-J, either, may not be theoretically and practically sufficient enough for the development of level (or proficiency) specific tests for each level of the scales.

Therefore, the primary purpose of this study is to examine the validity argument regarding the use of the CEFR-J for sound test development. A set of EFL tests were developed using the reading and listening scales, the procedure of which helps determine if the scale descriptors are sufficient enough for the development of level-specific EFL tests. Also, the test data helped empirically examine the nature of its developmental construct by looking at the two closely related levels of assessment at the item as well as the test levels. That is, the study empirically examined: 1) if the details of the CEFR-J reading and listening scales are sufficient enough for the development of level specific tests with the systematic increase of the mean difficulty from low to high levels, and 2) if the rank-ordering of the carefully constructed test items based on them is pertinent to their intended levels.

3. Method

Participants

For the reading part of the data collection, 412 freshmen (126 male; 286 female) took the reading tests. Their English proficiency greatly varied with their TOEFL ITP scores ranging from 370 to 550. They took two versions of the reading tests, Forms A and B, each of which consisted of 12 passages and 32 items. The tests included test items constructed targeting the five CEFR-J levels of A1.3, A2.1, A2.2, B1.1, and B1.2.

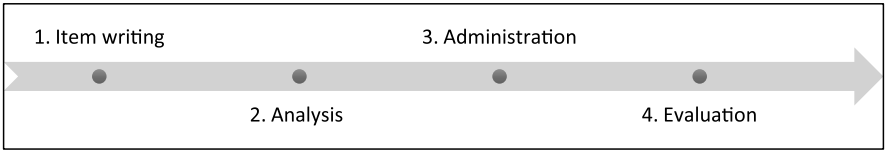
For the listening test data, 217 1st through 4th year students (81 male; 136

female) sat the listening exams. Their English proficiency also varied with the TOEFL ITP scores ranging from 370 to 600. The listening test also included two forms, Forms A and B; however, each form was administered to a different group of the students. Form A included 28 items, 17 of which were common items for test equation between the two forms. Form B also included the same number of common items and 12 unique items, making the total of 29 items. The listening test items were developed targeting the seven levels of A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, and B2.1.

Test development

The tests administered in this project were developed through four stages as Figure 1 demonstrates. A number of EFL instructors with experiences in EFL teaching developed the test items after going through a series of internalization processes of the CEFR-J reading and listening scale descriptors. Another group of researchers carefully checked the test materials developed by the EFL instructors for their pertinence to their intended levels. If the quality of test items and/or their source materials (e.g., reading passages and listening scripts) and their pertinence to the intended levels were in question, such items and/or their materials were either revised or simply abandoned.

Figure 1. Stages of test development



The final versions of the tests were piloted by having a small group of students with varying English proficiency. The data obtained from the piloting procedures helped examine the statistical quality of the individual test items. Test items that exhibited poor statistical quality were excluded from the tests. Multiple administrations of the tests yielded two sets of reading and

listening test data, which were subsequently analyzed to address the research questions.

Phases of test construction

Figure 2 demonstrates the developmental phases of the test instruments in more detail. The test writers were required to make sure that they follow the two phases of familiarization and source materials collection before they begin to write actual test items.

Figure 2. Phases of test construction

As Table 3 details, in the familiarization phase, they internalized the descriptors of the target CEFR-J reading and listening scales. Only when the test writers felt sufficiently comfortable with the level descriptors, they began to collect materials that correspond to the task features and text types of each level.

The collected source materials were examined before they were used for actual test development. A group of researchers evaluated text materials for their pertinence to the intended levels using an evaluation manual. They checked if the text materials fully represent the text type depicted in the CEFR-J scales including the right level of lexical items. All reading passages and listening scripts were evaluated on a 4-point scale of 1 (Strongly agree), 2 (Agree), 3 (Disagree) and 4 (Strongly disagree), and only those that received the rating of 1 were used for further test development. With the source materials approved by the researchers, the test developers created test items that reflect task features (or functions) at each level.

Table 3. Specifications of reading and listening test construction

	Scales	
Phases	**Reading**	**Listening**
Familiarization	• Target level: A1.3, A2.1, A2.2, B1.2, B1.2 (5 levels)	• Target levels: A1.2, A1.3, A2.1, A2.2, B1.1, B1.2 and B2.1 (7 levels)
Materials collection & evaluation	• 51 passages in total	• 33 scripts in total
Test construction	• 2–4 items per passage • 62 items (27 passages) • Two versions of the reading test	• 1–3 items per conversation and monologue • 43 items (26 conversations and monologues) • Two versions with 28 and 29 items each (17 common items)

Analyses

The test data were subjected to statistical analyses using classical test/item analyses and BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) to estimate item parameters and equate different forms of the tests. With the reading test data, the 3-parameter model was run and the common person equation was used so that the two test forms of A and B were concurrently equated to place the adjusted parameter values on the same logit continuum. With the listening test data, the one parameter logistic model was employed due to the limited sample size. Also, to examine if a predicated model of the target levels collaborates with their difficulty progression, Bayesian informative hypothesis testing was conducted using the Comparison of Means (Kuiper & Hoijtink, 2010).

4. Results

The test results from the two forms of the reading and listening tests were combined and their descriptive statistics are presented in Table 4.

Table 4. Descriptive statistics for each test set

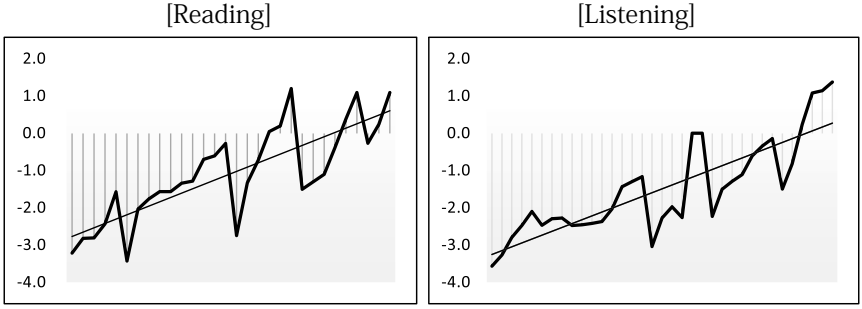
	Mean	Median	*SD*	*Empirical r*
Reading	20.9	21.0	3.82	0.76
Listening	19.7	20.5	4.29	0.74

The reading and listening test data demonstrate similar distributions with close central tendency while the listening shows more diversion (*SD*=4.29) compared to the reading test (*SD*=3.82). The rescaling procedure of the parameter values through a common item equation yielded empirical reliability coefficients of the entire tests, 0.76 and 0.74, respectively. The reading test resulted in a higher coefficient. The items that exhibited poor model fit indices were removed from the final test instruments; three items from the reading tests and five items from the listening tests were discarded from the final versions of the tests.

Difficulty progression by sub-level

Both of the reading and listening test data were reorganized using the logit values of the items under the same sub-levels and ordered them with the items of the lower levels on the left extreme and the items of the higher levels on the right as presented in Figure 3. As it reveals, both tests include some amount of deviations in their presentation of the consecutive increment of logit difficulty. The increase of the deviation from the expected trend is clearly noticeable around the mid-levels in both tests. Nonetheless, the linear trendlines across the sub-levels in both test data exhibit the increment of logit difficulty from lower to higher levels.

Figure 3. Logit difficulty rank-order by item



Bayesian hypothesis testing

Although the trendlines in Figure 3 demonstrated general progression of item difficulty in the intended and hence desirable direction, the amount of deviations some items exhibited create uncertainty as to the difficulty progression of the sublevels. That is, the mean logit values of the test items representing each sub-level (hence, sub-test) need to be evaluated for their progression of test difficulty. Therefore, Bayesian testing was performed with the reading and listening data. The test items were grouped together for each level and their mean logit values were examined using the Comparison of Means (Kuiper & Hoijtink, 2010). Bayesian hypothesis testing (Mackey & Ross, 2015) can help determine if the deviant pattern of the difficulty progression shown in Figure 3 can still be considered implicational with the level specific tests. The nonconformity of any individual levels can be tested against their predicted difficulty to see if such nonconformity could be ignored.

For the Bayesian procedures, the five sublevels were tested for their predicted implicational hypothesis; from A1.3 to B1.2 for the reading test and from A2.1 to B2.1 for the listening test. These five target sublevels were chosen to be tested as they resulted in with the most amount of deviation based on the item level analyses. Hence, it was examined if the mean difficulty at each level on each of the sampled tests increases symmetrically against the other four alternatives using Comparison of Means. The predicted hypothesis

was set as $\mu1 < \mu2 < \mu3 < \mu4 < \mu5$, which suggests that the levels present increasing difficulty from $\mu1$ to $\mu5$. The other four alternative hypotheses are as follows:

- H_0: $\mu1 = \mu2 = \mu3 = \mu4 = \mu5$
- H_a: $\mu1, \mu2, \mu3, \mu4, \mu5$
- H_2: $\mu1 = \mu2 < \mu3 = \mu4 < \mu5$
- H_3: $\mu1 < \mu2 = \mu3 < \mu4 = \mu5$

First, with the reading sublevel tests, the Bayes factor and the PMP were estimated, and the predicted hypothesis was compared against the other four alternatives using the values. Among the five hypotheses, the most supported one was the predicted hypothesis, with 46.04 of the Bayes factor and 0.39 of the PMP. Therefore, the implicational hypothesis is superior to the other hypotheses in terms of model-data fit. That is, this predicted hypothesis is empirically better supported by the data than the other hypotheses. Next, the listening tests were analyzed using the same Bayes procedures. Among the five hypotheses, the most supported one was the predicted hypothesis, with 32.74 of the Bayes factor and 0.53 of the PMP.

Therefore, for both reading and listening sublevel tests, the implicational hypothesis is superior to the other hypotheses in terms of model-data fit. In other words, the ordering of mean difficulties predicted by the specifications of the CEFR-J reading and listening scales is corroborated by the empirical data obtained from the examinee participants in the current study.

5. Discussion

To address its research purposes, this study examined the developmental processes of EFL reading and listening tests based on the CEFR-J scale descriptors. The rigor taken at the stages of test development and the statistical findings helped examine to what extent the developmental construct of the FL reading and listening depicted in the CEFR-J scales may be

considered valid.

A couple of issues were noted by the test developers with respects to the linguistic features of the scale descriptors. They found some level descriptors (e.g., A2.2 and B1.1) were not sufficient in their specificity for the text types and cognitive operations required for those levels. This finding is in line with the criticism often expressed by researchers (e.g., Weir, 2005; Fulcher, 2010). Also, the specifications frequently resort to degree words (e.g., *slow, slowly, clear, or clearly*) across adjacent levels. Especially in listening, the personalization of the listening stimuli (e.g., familiar to me) is common making it difficult to decide who the target learners should be with the test items. For example, B2.1 states that learners are able to read texts "within my field." However, it is not possible to know what field the learners would be in as the scales and the tests developed based on the scales are to serve general learner population. Another issue frequently commented by the test developers were about the length of the passages or scripts; how much longer should items in the C levels be than the B levels. Since there is no specification regarding the length, they had to depend on their own experiences and adjust the length considering the relative difficulty of adjacent levels, making the higher level passages simply longer than the lower ones. Apparently, if a test developer continues to follow this relative approach in deciding the difficulty of test items between two adjacent levels, the scores resulted from such tests will only be interpretable in subjective terms.

At the test level of statistical exploration, the results of the Bayesian testing procedures supported the predicted hypothesis of the five levels for both target scales of reading and listening. That is, despite the issues emerged at the stages of the test development, the empirical data collected from the administration of the tests suggested that the difficulty ordering of the tests was corroborated by the data obtained from the FL learners in this study. However, the model comparison technique of Bayesian testing is only to confirm that the hypothesized model be superior to other alternatives. That is, the procedure cannot completely rule out the possibility that the ordering of

the levels in question is not entirely implicational.

6. Conclusion

The study examined the use of the CEFR-J reading and listening scales for test development. It empirically tested the validity argument as to the CEFR-J reading and listening scales as a framework for FL test design. A set of tests were developed through rigorous procedures to assure their quality, and the rank-order of the test items were examined using their calibrated difficulty at the item as well as test levels. These procedures informed if the scale descriptors would lead to the development of level specific tests.

While the level specifications of the CEFR-J scales require much more specifics in realizing the developmental construct, the development of level-specific EFL reading and listening tests appears feasible as the items rank-ordered according to their difficulty logit values demonstrated a general progression from low to high levels. Bayesian testing procedures confirmed such a progression can be considered valid suggesting that developing level specific tests may be possible.

Finally, one needs to note the findings in this study suggest that the feasibility of a level-specific test based on the CEFR-J scales be possible only through a rigorous coordination of the test development procedures.

Acknowledgment

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

References

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380-390.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language testes: A matter of effect. *Language Teaching, 40*, 231-241.

Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR)

and effect-driven testing. *Advances in Research on Language Acquisition and Teaching: Selected Papers,* 15-26.

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

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods, 15*(1), 69-86.

Lantolf, J. P., & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*(2), 181-195.

Mackey, B. & Ross, S. J. (2015). Bayesian informative hypothesis testing. In L. Plonsky Ed.), *Advancing quantitative methods in second language research.* New York: Routledge.

Nagai, N., & O'Dwyer, F. (2012). The actual and potential impacts of the CEFR on language education in Japan, *Synergies Europe n° 6*, 141-52.

Negishi, M., Takada, T, & Tono, Y. (2012). A progress report on the development of the CEFR-J. *Studies in Language Testing, 36,* 137-165.

North, B. (1991). Standardisation of continuous assessment grades. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 167-177). London: Modern English Publication.

North, B., Martyniuk, W., & Panthier, J. (2010). Introduction: The manual for relating examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language Education. In Martyniuk, W. (Ed.), *Aligning Tests with the CEFR.* Cambridge: Cambridge University Press.

North, B., Ortega, A., & Sheehan, S. (2010). *A core inventory for general English,* British Council/ EAQUALS. Retrieved from http://www.teachingenglish.org.uk/sites/teacheng/files/Z243%20E&E%20EQUALS%20BROCHURErevised6.pdf.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

Pienemann, M., & Johnston, M. (1987). Factors affecting the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp.45-141). Adelaide: National Curriculum Resource Centre.

Runnels, J. (2013). Preliminary validation of A1 and A2 sub-levels of the CEFR-J. *Shiken Research Bulletin, 17*(1), 3-10.

Spolsky, B. (1986). A multiple choice for language testers. *Language Testing, 3*(2), 147-58.

Tono, Y. (2013). *CEFR-J Guidebook,* Tokyo: Taishukan Publishing.

Tono, Y., & Negishi, M. (2012). The CEFR-J: Adapting the CEFR for English language teaching

in Japan. *Framework & Language Portfolio Newsletter, 8,* 5-12.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Hampshire, UK: Palgrave-Macmillan.

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.