Development of Level-Specific Tests Using the CEFR-J Listening Descriptors

Development of Level-Specific Tests Using the CEFR-J Listening Descriptors

Megumi Sugita

Abstract

While the use of CEFR-J, a localized version of the CEFR for the Japanese contexts (Tono & Negishi, 2012), can provide firm bases for program development and test design, researchers (e.g., Fulchur, 2010; Runnels, 2013) have expressed concerns regarding its illustrative nature and the absence of an underlying psycholinguistic theory. In this study, English listening tests were developed based on the CEFR-J and were administered to 217 English learners. IRT item analyses and the Bayesian hypothesis testing were conducted to examine: 1) if the rank-ordering of the carefully constructed test items is pertinent to their intended levels of A1.2 through B2.1, and 2) the use of the CEFR-J listening scales helps to develop level specific tests with the systematic increase of the mean difficulty from low to high levels. The results indicate that the items rank-ordered based on their difficulty parameters demonstrated an implicational progression from A1.2 to B2.1; however, when the logit means of the sub-levels were considered, the distinction between A2.2 and B1.1 was not clear. Finally, it was confirmed that the development of a level-specific listening test based on the CEFR-J may be feasible when the development procedures are carefully coordinated.

I. Introduction

For the past few decades, foreign language (FL) proficiency scales and guidelines have been developed and gained popularity serving various educational purposes such as curriculum development, test design, and program evaluation (e.g., ACTFL Guidelines, CEFR, and Australian Second Language Proficiency Rating Scale). While some of them were employed for their intended uses in a limited educational context, others such as the ACTFL Guidelines and Common European Framework of Reference for Languages (CEFR) have been adopted or localized for the use in other educational contexts.

FL proficiency scales are usually intended to guide program development, instruction, and assessment. FL teachers may wish to employ such scales in their construction of tests or syllabuses concerning real-life tasks (North & Schneider, 1998). For example, the scales may help increase the reliability of subjectively judged ratings (Alderson, 1991) and also can provide guidelines for test construction (Dandonoli & Henning, 1990). Moreover, the scales can offer coherent internal links within an institution between pre-course testing, syllabus planning, materials organization, progress assessment and certification (North, 1991). Alternatively, they may help compare systems or populations using a common metric or yardstick (Liskin-Gasparro, 1984; Bachman & Savignon, 1986).

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

Despite the advantages above, numerous researchers have expressed concerns regarding (the use of) the scales (e.g., Spolsky, 1986; Pienemann & Johnston, 1987;

Alderson, 2007; Hulstijn, 2007; Runnels, 2013). As Lantolf and Frawley (1988) argue, one cannot simply assume that the progressive level distinctions and the number of levels are accurate, valid, or balanced; not to mention that the level specific descriptors are accurate, valid, or balanced (North & Schneider, 1998). Since FL scales are designed context-specific, the general use of them in a different context must be warned against (Spolsky, 1986).

Concerning CEFR and CEFR-J, researchers (e.g., Hulstijn, 2007; Runnels, 2013) have voiced concerns especially regarding their illustrative nature in describing learner performance in FL and the absence of an underlying psycholinguistic theory to explain the developmental construct of FL proficiency. The purpose of this study, therefore, is to examine the validity argument of CEFR-J, a localized version of CEFR, for its use as a framework for test design that can assess the proficiency development of English as an FL by Japanese learners of English in Japan.

II. Background

CEFR and CEFR-J

CEFR (Common European Framework of Reference for Languages), established by the Council of Europe, has been employed by a growing number of educational institutions today. It was designed to function as guidelines for all aspects of language teaching and learning including planning, instruction, and assessment. Its fundamental idea is based on "plurilingualism," in which individuals are expected to use different languages in different settings to interact with others. Another principle underlying CEFR is the action-oriented approach, which assumes language learners as acting socially using the target language. The framework thus provides descriptors with can-do statements in different levels of language competencies in reading, writing, listening and

speaking (interaction and production), and it consists of six levels from A1 (Basic) to C2 (Proficient) as shown in Table 1.

CEFR was originally developed in Europe where people are constantly traveling across national borders and are exposed to plural languages in their daily life. The framework rooted in the European setting, therefore, needed to be modified to the Japanese context, if the educators would like to apply it to the Japanese learners of English (Tono, 2013).

In 2004, the Koike Grant-in-Aid for a Scientific Research Group initiated the development of CEFR-J, which is a localized version of CEFR, and later the Tono Group took over the project. While CEFR sets A1 as the lowest level, the Tono Group decided to create another level (Pre-A1) below A1. It was due to the fact that most of the Japanese learners fall on Level A or have not even reached A1, and the descriptors in A1 in CEFR do not precisely describe what the Japanese learners are actually able to do using English (Negishi, 2012). In the same manner, A1 level was divided into three sub-levels (A1.1, A1.2 and A1.3) in CEFR-J, and A2, B1, and B2 into two levels (as shown in Table 1.). No change was made to C1 and C2.

Table 1. 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	
Proficiency user	C1	C1		
	C2	C2		

Issues concerning CEFR and CEFR-J

As mentioned earlier, while other FL proficiency scales suffer from their own empirical as well as theoretical issues, CEFR (and the CEFR-J) have been criticized mainly because of their illustrative nature in describing L2 development. The can-do statements that CEFR (and CEFR-J) employ exemplify what FL learners should be able to perform to be qualified for the intended levels. However, the statements only characterize what the learners can do, but not to what extent they can perform a given task. CEFR's (and CEFR-J's) illustrative nature was therefore often the reason for criticism as the can-do statements would not help test design and evaluation (Weir, 2005).

Likewise, the definitional vagueness of can-do's entails additional concerns regarding how L2 learners' performance on a given task should be interpreted in its completeness. Can-do is indeed not an absolute terminology and reasonably subjective within and across individual L2 users in their judgment of the target language performance.

Another issue with CEFR (and CEFR-J) comes from the absence of an underlying psycholinguistic theory. This leads to the lack of the evaluative means for the validation of the developmental FL construct. As CEFR (and CEFR-J) does not present a theoretically driven construct definition of FL proficiency, it is difficult to interpret what it means for a learner to know how to perform an FL task in the developmental perspectives.

Concerning CEFR-J alone, the addition of sub-levels not only intensifies the concerns regarding its use for test design and performance interpretation but also increases difficulty in distinguishing between adjacent levels around the sub-levels of A and B. For instance, the test developer in this study reported that she had to deal with a number of specification confusions concerning text types and operations as illustrated in terms of can-do's especially within and across the levels of A1.3 through B1.2.

As a result, CEFR's (and CEFR-J's) no theory-bound approach to its definition of proficiency construct has led its users to understand the scales as a rather heuristic model (Fulcher, 2010). To serve as a common reference for assessment, both CEFR and CEFR-J would require empirical evidence for the very nature of the developmental construct, particularly, of FL listening, the FL skills that have seen only a limited scope of the investigation until now. That is, the developmental construct of FL listening needs to be empirically evidenced and should be demonstrated with reference to the rank-ordering of the carefully constructed test items that are pertinent to the descriptors of the CEFR-J listening scales.

III. Purpose

For the demonstration of the validity argument regarding the use of CEFR-J for sound test development, the nature of its developmental construct needs to be empirically evidenced at the two closely related levels: 1) the item level, and 2) the test level. At the item level, the rank-ordering of individual test items needs to be evidenced with their calibrated item difficulty from the lower to the higher CEFR-J sub-levels. At the test level, the systematic increase of the mean difficulty needs to be demonstrated from the lower to the higher test levels. Moreover, the test items intended for the same level should work together to form a level specific (sub-)test representing their intended (mean) difficulty.

Therefore, the purpose of this study is to develop a set of EFL listening tests using the level specifications of the CEFR-J listening scales and validate if such a use of the scales helps design psychometrically sound level specific tests. In order to achieve the research purpose, this study examined the following two questions systematically:

1) if the rank-ordering of the carefully constructed test items in this study is

pertinent to their intended levels of A1.2 through B2.1; and

2) if the use of the CEFR-J listening scales helps to develop level specific tests with the systematic increase of the mean difficulty from low to high levels

IV. Methods

Participants

The test data used for this study come from 217 English majors at a university in Japan. Their school years vary from first to fourth years; 136 students were female and the rest 81 male. The level of the students' English proficiency varied greatly with most of their TOEFL ITP scores falling between 370 and 587 at the time of data collection.

Test instruments

The entire process of test development included four stages as depicted in Figure 1. A professional EFL instructor with expert knowledge in test development was hired, and developed two sets of EFL listening tests based on the descriptors of the CEFR-J listening scales. Then, the researchers examined the pertinence of the texts and the items to their intended levels, and when the pertinence was in question, the texts and their items were either revised and included in the item pools or abandoned.

Figure 1. Stages of test development

Upon completion of the final versions of the test instruments, the researchers

administered them to a group of students with varying English proficiency. Multiple administrations of the tests yielded the test data, which were subsequently entered to the analyses to address the research questions.

Phases of test construction

As Figure 2 exhibits, in developing the initial versions of the test instruments, the test writer ensured the two phases of *familiarization* and *source materials collection* before she began to write actual test materials. In the familiarization phase, she internalized the descriptors of the seven target CEFR-J listening scales; A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, and B2.1.

Figure 2. Phases of test construction

Whenever the test writer noticed a source of confusion in the descriptors, she inquired the researchers and documented all the challenges that she had to resolve while internalizing the descriptors of the target levels.

When the test writer felt sure about every aspect of the descriptors, she started to collect source materials that reflect task features and text types in each level descriptor. Many of the materials came from the Internet sources, while some of them came directly from the surroundings in her daily life, such as announcements on the train or at the airport and an interview on TV.

In Phase 3 (Test development), the test writer began to create test items to elicit the listening functions dictated in each level descriptor. She wrote one to three test items for each spoken text in the form of either a conversation or a monologue. She developed a

total of 43 test items based on 26 conversations and monologues, and the items were used to prepare two test forms; one form with 28 items and the other with 29 items. They shared 17 common items for subsequent test equation. These common items were used to estimate the item parameters of different tests and place them on the common logit scale of difficulty.

Analyses

The test data were statistically analyzed to determine the quality of the test items and the tests as a whole. In addition to classical test/item analyses, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) performed one parameter logistic model. One parameter model was employed due to the limited sample size. While BILOG-MG calibrated item parameters, the two test forms of A and B were concurrently equated to place the adjusted parameter values on the same logit continuum.

Bayesian informative hypothesis testing was also performed using the Comparison of Means (Kuiper & Hoijtink, 2010). The procedure tested if a predicted model of the five target levels results in with increasing difficulty.

V. Results

Tables 2 and 3 present the descriptive statistics for each test form.

Table 2. Descriptive statistics for FORM A (k=28)

Mean	Median	*SD*	Kurt.	Skew.	Range	*A*
20.5	21	3.75	-0.71	-0.17	10-27	0.71

Table 3. Descriptive statistics for FORM B (k=29)

Mean	Median	*SD*	Kurt.	Skew.	Range	*A*
18.9	20	4.61	-0.80	0.18	5-26	0.72

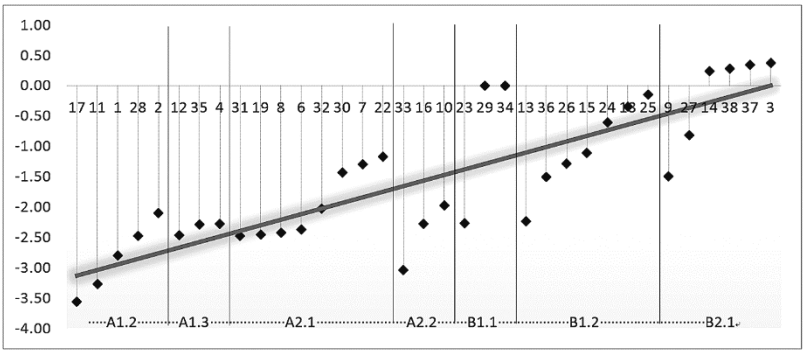
The data appear normal in their distributions considering their normality indices, such as kurtosis and skewness. Also, considering the mean values, Form B appears to have been easier than Form A. However, the reliability coefficients (i.e., Cronbach's alpha) resulted in at the lower limit of the acceptable range for both tests, 0.71 for Form A and 0.72 for Form B, respectively.

The rescaling procedure of the parameter values through a common item equation yielded an empirical reliability of the entire test, 0.74. It is not surprising to find a higher reliability, as it comes from a test of the two forms combined. In order to achieve an even higher test reliability coefficient, five items with poor model fit indices were removed from the final test instruments, and the subsequent analyses were conducted as such.

Difficulty progression by item

Using the difficulty parameter values produced by BILOG-MG, the test items are rank-ordered across the adjacent sub-levels, as shown in Figure 3. The trendline that goes through the logit points exhibits an implicational progression with consecutive difficulty increments of test items from the lower to the higher sub-levels.

Figure 3. Difficulty rank-order by item



While the trendline demonstrates the increasing difficulty of the items across the sub-levels, there are a few items that fall away from the trendline. The deviance of these test items may have occurred for several reasons. The items may not have correctly represented the specifications of their intended levels; i.e., their pertinence to the intended levels was not (fully) met. The level specifications may not have been sufficient enough to guide the development of the items with an appropriate level of difficulty. Alternatively, while the first two conditions were satisfied, the specifications themselves were not theoretically sound in their presentation of the FL listening development, and hence, it was not possible to correctly realize the developmental construct of FL listening in the test items.

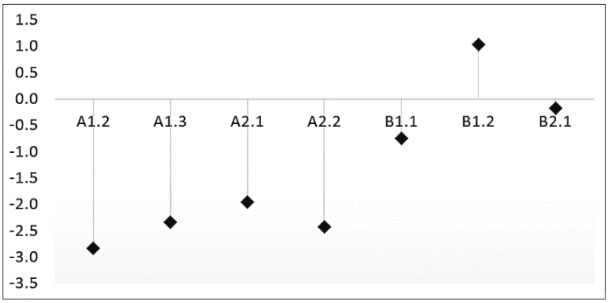
Considering the amount of rigor that the current study invested in test development, however, it is unlikely that the test items did not adequately reflect the specifications. Therefore, the possibility appears to lie with the characteristics of the specifications themselves. That is, as often criticized by other researchers (e.g., Weir, 2005), the specifications may not have been specific enough for test development. Alternatively, due to the lack of specificity as to how FL listening develops (Fulcher, 2010), text types may not be correctly ordered and therefore erroneously designated in the scales. Likewise, the listening functions were not correctly ranked and designated accordingly as the FL listening functions exemplified in the scales are neither theoretically motivated nor empirically validated as to their implicational order.

Especially, as Figure 3 exhibits, the items for A1.2 through B1.2 fall below 0 of the logit. That is, the specifications may not be specific enough to address the linguistic as well as cognitive challenges projected for the upper levels. Consequently, the overall difficulty of the entire test resulted in easier than it was supposed to be considering the examinees' English proficiency of the low to the intermediate.

Difficulty progression by sub-level

The data were reorganized using the mean logit difficulty of the items under the same sub-levels and rank-ordered them, as presented in Figure 4. As it reveals, the mid-levels do not demonstrate the consecutive increment of logit difficulty. Especially, A2.2 and B2.1 resulted in easier or B1.2 more difficult compared to their adjacent levels.

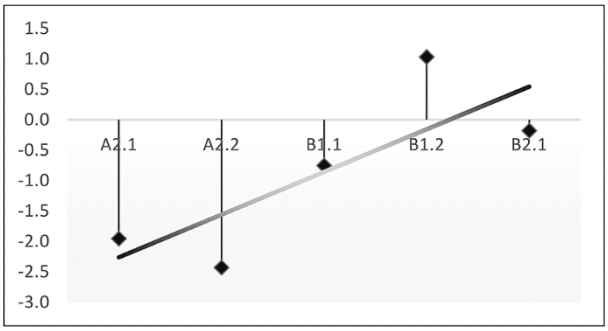
Figure 4. Difficulty rank-order by level



Bayesian hypothesis testing

For a closer look at the reversed ordering of the sub-levels, the five sub-levels A2.1, A2.2, B1.1, B1.2, and B2.1 were further examined using Bayesian hypothesis testing (Mackey & Ross, 2015). Figure 5 highlights the unordered difficulty progression of the five levels, A2.1 through B2.1. Although the trendline demonstrates a general progression of increasing difficulty, the average logits of A2.2, B1.2, and B2.1 do not stand close to the line.

Figure 5. Logit difficulties of the five sub-levels for Bayesian testing



Bayesian hypothesis testing was performed to check if the deviant pattern of the difficulty progression shown in Figure 5 can still be considered implicational, and the nonconformity of A2.2, B1.2, and B2.1 in contrast to their predicted difficulty could be ignored at least mathematically. The Bayesian procedures, hence, tested the predicted implicational hypothesis (i.e., if the mean difficulty at each level on each of the sampled tests increases symmetrically from A2.1 to B2.1) against the other four alternatives using Comparison of Means (Kuiper & Hoijtink, 2010). 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: $\mu1 < \mu2 = \mu3 < \mu4 = \mu5$

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 6.14 of the Bayes factor and 0.53 of the PMP. Therefore, the implicational hypothesis, at least mathematically, 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 collapse-down hypothesis H_2, the Bayes factor and the PMP of H_2 of which were 3.85 and 0.33, respectively. In other words, the ordering of mean difficulties predicted by the specifications of the CEFR-J listening scales is corroborated by the empirical data based on the examinee participants in the current study.

VI. Discussion

The statistical properties of the test data helped address several issues concerning the developmental construct of the FL listening depicted in the CEFR-J listening scales. They also helped explore the possibility as to the development of level-specific English listening tests.

While the rank-ordered items according to their difficulty parameters demonstrated a general progression from A1.2 to B2.1, some of them were not observed around their predicted difficulty represented as the trendline in Figure 3. Notably, the test items for A2.2 and B1.1 were not corroborative to their intended levels in concert, as Figure 3 reveals, and the cause of such disconformity was pursued by examining their specifications in Table 4 closely.

Table 4. The listening specifications of A2.2 and B1.1 (from CEFR-J1.0Eng)

A2.2	I can understand and follow a series of instructions for sports, cooking, etc. provided they are delivered slowly and clearly.
	I can understand instructions about procedures (e.g., cooking, handicrafts), with visual aids, provided they are delivered in slow and clear speech involving rephrasing and repetition.
B1.1	I can understand the gist of explanations of cultural practices and customs that are unfamiliar to me, provided they are delivered in slow and clear speech involving rephrasing and repetition.
	I can understand the main points of extended discussions around me, provided speech is clearly articulated and in a familiar accent.

As often criticized by researchers (e.g., Weir, 2005; Fulcher, 2010), the specifications of A2.2 and B1.1 appear problematic, as they significantly lacked in specificity for the text types and the operations of FL listening. In particular, the specifications include a number of degree words such as *slow, slowly, clear*, and *clearly* and also resort to the personalization of the listening stimuli, e.g., *unfamiliar to me, around me,* or *a familiar accent*. The subjectivity that these terminologies imply only makes the development of test items for these two levels difficult, especially, in relation to the adjacent levels. That is, the two levels would not stand as independent, and the specifications would be difficult to interpret and hence will not help develop level specific tests of their own.

Moreover, the developed test items as a whole were found to be much easier than they were supposed to be. Except the four items for B2.1, the logit values of the other sub-levels all fell below 0. This overall easiness of the test items may be due to the level of lexical items employed in the listening stimuli of conversations and monologues. Also, the length of the speech, whether it be a conversation or a monologue, may have also been an issue, as it significantly affects the item difficulty. Since the length of the stimuli often depends on the text types (e.g., announcement or assignment), many of the scripts

were short. Also, for the memory not to be an issue in listening, the test writer made efforts to keep listening stimuli short. Those factors may have affected the overall difficulty of the tests in this study.

At the test level exploration, the results of the Bayesian testing procedures supported the predicted hypothesis of the five levels from A2.1 to B2.1 most. Therefore, the difficulty ordering of the tests developed based on the specifications of the five levels was corroborated by the empirical 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 the other alternatives. That is, the procedure cannot completely rule out the possibility that the ordering of the levels in question is not entirely implicational.

In sum, considering all the statistical results combined, the development of level-specific FL listening tests may be feasible using the CEFR-J listening scales of A1.2 through B2.1. The entire development procedures, however, should be rigorously supervised as this study demonstrated.

VII. Conclusion

The purpose of this study was to explore the use of the CEFR-J listening scale for test development; i.e., the validity argument as to the CEFR-J listening scales as a framework for FL test design. The test items were rank-ordered according to their calibrated difficulty and were examined for their pertinence to their intended levels. The increments of the average difficulty by level were also checked to see if the scale descriptors would enable the development of level specific tests.

This study found that while the level specifications of the CEFR-J listening scales require much more specifics in realizing the developmental construct, the development of level-specific FL listening tests appears feasible. The rigorous test development

procedures helped achieve test items of varying difficulty, and most of them were found pertinent to their intended levels. Such pertinence also helped develop level specific tests as the items worked in concert in generating mean logit values that were corroborative with the progression of the sub-levels.

While this study helped evidence and confirm several essential aspects of (the use of) the CEFR-J listening scales for test development, a couple of limitations need to be recognized. First, some levels (e.g., A2.2 and B1.1) were examined only with a limited number of items. As the test items at these levels were especially problematic, a future study should examine the levels with more items. Second, this study only examined the CEFR-J levels up to B2.1 due to the limited English proficiency of the examinee population. In order to examine the developmental construct of the entire CEFR-J scales, the study should have included test items representing the upper levels and also examinees of high English proficiency.

Acknowledgment

I would like to thank Professor Tetsuko Fukawa, Professor Yasuko Ito, and Professor Siwon Park for their invaluable 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.

VIII. References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71-86). London: Modern English Publication.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659-663.

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.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals*, *23*(1), 11-22.

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.

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: A historical perspective. In T. V. Higgs (Ed.), *Teaching for proficiency, the organizing principle* (pp. 11-42). Lincolnwood, IL: National Textbook Company.

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

Negishi, M. (2012). *The Development of the CEFR-J: Where We Are, Where We Are Going*. Project Report on Grant-in-Aid for a Scientific Research (KAKENHI).

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., & 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.