CEFR-J Based Survey for Japanese University Students

Masanori Tokeshi, Meio University, Japan
Lianli Gao, Academy of Chinese Medicine Sciences, China

The European Conference on Education 2015
Official Conference Proceedings

Abstract

This study attempts to empirically examine self-ratings of Can-Do descriptors of the CEFR-J (Tono, 2012), which was modified from the CEFR (Council of Europe, 2001). The study mainly focuses on the relationship between the English proficiency test (EIKEN) scores and self-ratings and reliability of self-ratings between five skill categories. Three hundred eighty-nine freshmen at one Japanese university answered a web-questionnaire (110 questions in five skill categories) based on the CEFR-J Can-Do descriptors. The results show contradictory evidence. According to an in-depth investigation of individual raw data, the results indicate a variation of responses with little relation to English proficiency test scores. A statistical analysis (Pearson's R) also supported this evidence. However, the results also indicate that the internal reliability of self-ratings between the five skill categories is high, according to Cronbach's alpha value (0.872), when the data were compared in the group. To interpret this contradictory evidence, it may be inferred that CEFR-J is effective to evaluate general proficiency skill levels of overall English programs, but not very helpful to measure individual English learning.

Keywords: CEFR, assessment, curriculum, higher education, Japan

iafor
The International Academic Forum
www.iafor.org

1. The background to Japanese university English education

The Japanese Ministry of Education proposed the English education reform (2014) by urging all levels of schools to strengthen English education, especially in order to make rapid preparations for the Tokyo Olympic Games in 2020. The strong rationale for implementing the English education reform is partly due to the fact that the average score (70 points) of Japanese examinees in TOEFL iBT ranked 25th out of 30 Asian countries, whereas rival neighboring area/countries' examinees gained much higher average scores, respectively '77' for China, '85' for Republic of Korea (South Korea) and '79' for Taiwan (ETS, 2014).

It is essential for Japanese universities to set English proficiency standards which enable students to equally compete with those around the world. However, "we have not had any agreed attainment targets in language teaching so far. Nor do we have any consensus as to how to attain those targets or how to assess the attainment" (Negishi, Takada, & Tono, 2012, p.136) at university level schools in Japan. That's partly because there is no core university English curriculum specified by the Ministry of Education. Currently, TOEIC, Test in Practical English Proficiency (hereafter, EIKEN test) or GTEC (Global Test of English Communication) have often been used to assess English proficiency of university students. These three tests assess mainly listening and reading, although they have been developing new types of tests measuring four skills. TOEFL iBT comprehensively measures four skills and is taken in the largest number of countries (roughly 130 countries) and adopted in educational institutions (roughly 9,000) in the world, which makes it the strongest candidate for test takers, institutions and teachers worldwide to compare English scores across country/area. However, there are practicality problems such as expensive testing fee, availability of testing centers and trained raters (Tokeshi, 2013). Therefore, it is hard to claim that any specific major commercial English test can be exclusively used by all university teachers in Japan to confirm attainment in English proficiency of their students. As globalization spreads rapidly, it is extremely important for not only university educators, but also students to understand whether English proficiency of students gained through English curricula or self-learning is high enough to be able to compete with students from other countries. If English proficiency of our students is not equal to or lower than that of students from other countries, it is also important to know what type of teaching and learning should be included in the English curriculum. More appropriate information is needed to communicate about English proficiency of our students among educators, students and stake-holders.

The Common European Framework of Reference for Languages (hereafter, CEFR) (Council of Europe, 2001) seems the most promising English education toolkit (North, 2014) and it has become the international standard for language teaching and learning (North, Ortega, & Sheehan, 2010). Figueras (2012) states the CEFR would be the

most relevant and controversial document in language teaching, learning and assessment fields in the twenty-first century. Alderson (2007) claims that despite criticism from some researchers, "nobody engaged in language education in Europe can ignore the existence of the CEFR"(p.660).
The CEFF-J (Tono et al, 2012), which this study attempts to validate, is the Japanese version of the CEFR and it includes twelve levels of English proficiency specifications designed by two research groups of the Ministry of Education Grant-in Aid projects over eight years (from 2004 to 2011), which will be discussed in detail subsequently. It is designed to be the most tailored to the Japanese context (Negishi, Takada, & Tono, 2012).

The CEFR-J still remains to be empirically examined before being adopted in any university or other level of schools in Japan. Despite longitudinal and elaborate research projects, few studies (e.g., Runnel, 2013; Runnel, 2014) at university level have been conducted to examine the CEFR-J. "In fact, little research on the relationship between ability, self-assessment, and CEFR-aligned task performance for Japanese learners has been carried out" (Runnels, 2014, p.86).

It is urgently needed to validate the CEFR-J before being incorporated in the curriculum development, teaching and learning for university English education.

This study conducted a web-questionnaire for M University freshmen (N: 389) using 110 Can-Do statements adopted from the CEFR-J and seeks to examine the following research questions.

(1) Is there a relationship between self-ratings of Can-Do descriptors of five skill categories?
(2) Is there a correlation between EIKEN English Proficiency test scores and CEFR-J self-rating results?
(3) What implications for university English programs learned from the study?

2.0 Literature Review

2.1 The CEFR

The Council of Europe published the CEFR; Common European Framework of Reference for Languages (2001) after 30 year's research in Europe. Its publication dates back to communication/function related research by van-Ek (1975) and Wilkins (1976) in the 1970s and major features are based on earlier Threshold-series publications; "Threshold" (van Ek & Trim, 2001b), "Waystage" (van Ek & Trim, 2001a) and "Vantage" (van Ek & Trim, 2001c) published by the Council of Europe. The CEFR was developed based on two projects, DIALANG in 1996 (available in Council of Europe, 2001, p. 226-243) and the ALTE 'can-do' project (ALTE, 2002). Its principles reflect 'plurilingualism' and 'pluriculturalism' in the European context. The CEFR provides six levels (A1 to C2) of illustrative descriptors in five skill categories in which speaking is divided into spoken interaction and spoken production, in addition to listening, reading, and writing. It includes four domains of language use; public, personal, educational, and professional, for each of which locations, institutions, persons, objects, events, operations and texts are specified (Council of Europe, 2001, pp. 48-49).

The CEFR has some salient features of its strengths. The CEFR can be helpful as it helps to understand what is assessed, how performance is interpreted and how comparison across different tests and examinations can be made (Council of Europe, 2001). It is an action/outcome-oriented approach and a learner's language performance is calibrated against its standards. The Framework provides a self-assessment grip (Council of Europe, 2001, p.26) with a form of Can-Do statements in which a learner judges his/her own language ability as to what he/she 'can do' in a foreign language. Its focus is on communication and learner/user rather than on linguistic competence . It was developed to provide "a common basis for the explicit descriptions of objectives, content and methods" (Council of Europe, 2001, p.1) and expected to help develop course curricula, textbooks, and examinations.

Despite widespread use worldwide, researchers criticize some limitations of the CEFR. There is a mismatch between the influence of the CEFR and its adoption into curricula, pedagogy and assessment (Figueras, 2012).

One strong claim is its adoption for testing. Some testing researchers (Weir, 2005; Alderson, 2007; Little, 2010) are critical of its theoretical underpinnings for testing so that they strongly ask for empirical validation of it. "It is not surprising that a number of studies have experienced difficulty in attempting to use the CEFR for test development and comparative purposes" (Weir, 2005).

Jones's study (2002) is fairly relevant to this current study. Jones compared 'Can Do' self-ratings (questionnaire) with Cambridge examinations (KET, PET, FCE, CAE, CPE). The results showed that there was a great variation of perceptions on personal own language ability at the individual level. Interestingly, lower level of respondents tended to rate themselves too generously (higher than actual ability) and high level of respondents tended to rate themselves more modestly (lower than actual ability). He concluded that "people tend to understand 'can-do' differently" (p.33), depending on personal background such as age, first language and proficiency level. He assumed that "the problem is probably a particular feature of the present data, based on self-report" (p.33). Little (2010, pp.159-160) also points out concerns about self-assessment; 1) learners do not know how to assess themselves; 2) there is a danger that they will overestimate their proficiency; and 3) they may be tempted to cheat by including in their ELPs (hereafter, European Language Portfolio) material that is not their own.

Another issue related to this study is empirical validation of the CEFR. North (2014) claims that CEFR descriptors are scaled based on teacher's perceptions of the second language proficiency of learners. The descriptors have not emerged from in-depth, large-scale longitudinal studies of the actual process of second language acquisition over time (p.23). In line with issue of empirical validation, Hulstijin (2007) claims that qualitative and quantitative dimensions of language proficiency in the CEFR should be sufficiently validated by empirical studies (2007).

2.2 CEFR-J projects

Carrying over the previous project led by Koike (2004-2007), a new Grant-in Aid Scientific Research led by Tono (2008-2011) published 12 levels of the CEFR-J Version 1 (2012) to publicize the final result of the project. Research on the implementation of the CEFR-J began in 2008 at the Tokyo University of Foreign Studies. The project was carried out by a group of 18 researchers engaged in English education.

The CEFR-J projects were chronologically completed with the following six stages.

STAGE 1 (Y2006); A Can Do questionnaire was developed from DIALANG self-assessment statements (Council of Europe, 2001). It was translated into Japanese and the questionnaire was given to 360 Japanese university students (can or cannot dichotomy questions). Seven hundred twenty-seven Japanese upper secondary school and university students were investigated by using the same Can-Do descriptors accompanied by examples with four scales of answers. The results confirmed that the CEFR could be adapted to Japanese learners of English.

STAGE 2 (Y2004-2007); Various research was conducted to investigate English proficiency of the participants for different school levels of students (354 elementary schools, 150 junior & senior high schools) and for 7,354 business persons. Following the results, it was concluded that over 80% of English language learners in Japan fell within the A1 & A2 levels of the CEFR (also known as the Basic User level).

STAGE 3 (Y2008); Accordingly, the original six levels of the CEFR were divided into 12 levels for the Japanese version of the CEFR (CEFR-J alpha version). The alpha version of the CEFR-J was designed by considering ELP, Can-Do descriptors, GTEC tests, Super English Language-high schools, EIKEN tests. The special features of the CEFR-J are as follows (Negishi, Takada & Tono, 2012, p.143); 1) Add Pre-A1, 2) Divide A1 into three levels: A1.1, A1.2, A1.3, 3) Divide A2 into two levels: A2.1, A2.2., 4) Divide B1 into two levels: B1.1, B1.2., 5) Divide B2 into two levels: B2.1., B2.2., 6) No change for C1 and C2.

STAGE 4 (Y2009); After receiving some advice from a CEFR specialist, Dr. Anthony Green, productive skills were broken down into (1) performance, (2) criteria, and (3) condition, while those for receptive skills were broken into (1) task, (2) text, and (3) condition. Furthermore, the descriptors of the alpha version were sorted by 206 English teachers to ensure the appropriate order of difficulty and then were reordered according to the teacher survey. The orders were changed only when over 70% of the participating teachers agreed with the order of the descriptors. Thus, the CEFR-J alpha version was modified and the beta version of the CEFR-J was finalized.

STAGE 5 (Y2010-2011); To validate the beta version, 1,685 junior high school students, 2,538 senior high school students and 1,245 university students answered the questionnaire with four answer choices as to the degree with which they could do about all the descriptors in the questionnaire. To solve the problems identified in the statistical analysis of the beta version, the descriptor statements were modified and the order was changed again. Also, the project group implemented performance tests based on the descriptors for five skill categories in order to analyze the relationship between their self-assessment and their actual performance (Negishi, Takada & Tono, 2012, p.145).

STAGE 6 (2012-2013), Completing the validation processes, the CEFR-J Version 1 was released in March, 2012 (http://www.cefr-j.org/download.html) and the “CEFR-J Guidebook” was published in 2013.

Following publication of the CEFR-J, little empirical research had been done. However, Runnel's study (2014) investigated 590 Japanese university students. Her research results indicate that unfamiliarity and confusing content of can-do statements affected reliability of the hierarchy of the statements and individual differences in a population of the learners affected the results of difficulty of self-rating. The conclusion of her study requested further studies on the CEFR; "the CEFR-J's target users' responses to can-do statements, and content analyses of the can-do statements should be performed to ensure a consistent, common interpretation of the system" (p.86).

3. Method

3.1 Participants

Four hundred eighty-eight university freshmen at M University in Japan were asked to answer the web-questionnaire and 389 students answered the questionnaire. They were enrolled in 17 freshman English classes from three different departments, taught by 10 teachers in the first semester of 2014, when the questionnaire was given. The classes at that school were divided according to placement test scores before the classes began. M University was a public school and was selected since the freshmen at M University usually gained almost the average score of all examinees in the National Center for Entrance Examination, which the majority of high school students nationwide take in Japan. The students at M University were considered to represent the average English learners in the freshman year at university level in Japan.

The participants were only limited to those who agreed to answer the questionnaire. So, 389 participants from the target population (N: 488) participated in this project. Best efforts were made not to violate the participants' privacy. M University research grant committee gave the researcher permission. The researcher gained permission by email from the CEFR-J project team to download the CEFR-J Version.1 Can-Do descriptors from their homepage.

3.2 Research instruments

This study adopted a self-designed web-questionnaire written in participants' first language. The questionnaire used Can-Do descriptors available in the homepage of CEFR-J Version 1.1 (http://www.cefr-j.org/download.html) and included 110 can-do descriptors, respectively 22 questions for five skill categories (listening, reading, spoken interaction, spoken production and writing). The participants were asked to rate their own English ability for each Can-Do descriptor according to a scale (strongly disagree, moderately disagree, moderately agree, strongly agree).

Example of listening descriptors: Q7 (A1.3 level)
"I can catch concrete information (e.g. places and times) on familiar topics encountered in everyday life, provided it is delivered in slow and clear speech"

3.3 Data collection method

The pilot studies before the main web-questionnaire were conducted twice for 23 junior student taking 'English Teaching Methodology' class taught by the researcher. According to their feedback, the questionnaire was revised and with the help of the researcher's colleague, the web-questionnaire was designed and uploaded on the webpage in late April, 2014. The researcher asked the freshman English teachers to cooperate on the project. With their cooperation, between late April, 2014 and late May, 2014, for about a month, the participated students were asked to answer the web-questionnaire out of class with their cellular phone (QR Code) or with computer (URL).

As the research proposal admitted by the university research grant committee indicates, this study used EIKEN English Proficiency test (Type B) which is designed to assess a range of levels from EIKEN Grade 3 and EIKEN Grade 2. Most of the participants in this study (except five students) took this test for class placement purposes in early April, 2014.

3.4 Data analysis method

The web-questionnaire results were collected from the web-page and saved in EXCEL file and invalid participants' responses were excluded (ex., six participants gave the same ratings on all descriptors). The questionnaire results were compared with the English proficiency test results to check the correlation between them by using SPSS Version 21. In addition, an in-depth analysis of questionnaire results was conducted to see the relationship between the EIKEN test scores and the participants' self-ratings in the five skill categories.

4. Results and analysis

This section mainly discusses the results and its analysis regarding the relationship between the English proficiency test result (EIKEN score) and self-ratings as well as the relationship among Can-Do descriptors in the five skill categories.

First, numbers were substituted for the questionnaire responses to conduct quantitative analysis. In the following graph/tables, substitution for responses is; 'strongly disagree' is '1', 'moderately disagree' is '2', 'moderately agree' is '3', 'strongly agree' is '4'. The average self-rating of each skill category (e.g., listening) for individual respondents was calculated to see the relationship among the self-ratings for five skill categories. For example, the average self-rating of the EIKEN score 1st ranked respondent for 22 listening descriptors is '3.64' (see Table 1 below).

Table 1: CEFR-J levels and questionnaire questions (Qs)

Level:	PreA1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	C1	C2
Qs:	1&2	3&4	5&6	7&8	9&10	11&12	13&14	15&16	17&18	19&20	21	22

4.1 Relationship between self-ratings of CEFR-J descriptors and EIKEN test scores

Person's R (two sides) was utilized to examine the correlation between the English proficiency test, EIKEN scores and self-ratings for the five skill categories. As shown in Table 2, Person's R values for each skill category are low. The results indicate that the correlation between EIKEN test scores and each skill category has a weak relationship, respectively.

Table 2: Correlation between EIKEN scores and self-ratings of five skill categories

	listening	reading	spoken interaction	spoken production	writing
Pearson's R	.271	.292	.251	.312	.292

Furthermore, questionnaire raw data of individual respondents were examined to get in-depth analysis of the correlation between EIKEN test scores and self-ratings.

Due to limited paper space, 13 respondents' responses for listening descriptors were selected from every 30th rank (1st, 30th,60th ...360th) ordered according to EIKEN scores (see Table 3). This study adopted self-rating '4 (strongly agree)' squared with boldfaced lines in the table as the borderline of achieving the level (estimated 80% of achievement), following the criteria suggested by North (2014, p.103), stating "When a learner met 80% of the descriptors on the checklist for the level concerned, they could be considered to 'be' that level." When there is no clear cut-off between '4' point and '3' point or other points in the responses or when there is no '4' point, the lowest level of descriptor which has '3' was chosen. For example, for the 60th respondent, the lowest level of descriptor, Q1 was chosen as the borderline because there is no '4' in the responses (see Table 3).

Table 3: Raw data from Can-Do listening descriptor responses (a sample of every 30th rank according to EIKEN scores): bold-faced squares are borderlines

Level	Pre A1		A1.1		A1.2		A1.3		A2.1		A2.2	
ranking	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
1	4	4	4	4	4	4	4	4	4	4	4	4
30	4	4	4	4	4	4	4	4	3	3	3	3
60	**3**	2	3	3	3	3	2	3	2	3	3	3
90	**3**	3	3	3	3	3	3	2	3	3	3	3
120	4	3	4	4	4	**4**	3	3	3	3	3	3
150	**3**	3	3	3	3	3	3	3	3	3	2	3
180	**3**	3	3	3	3	2	3	3	3	3	3	3
210	**3**	2	3	3	3	3	3	3	3	3	2	3
240	4	3	4	4	4	4	3	4	3	4	4	**4**
270	3	3	4	4	3	3	3	**4**	3	3	3	3
300	**3**	3	4	3	3	3	3	2	2	3	2	2
330	**3**	2	3	3	2	2	2	2	1	3	2	2
360	**3**	2	4	3	3	3	2	1	3	3	3	3

(continued)

Level	B1.1		B1.2		B2.1		B2.2		C1	C2	
ranking	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	mean
1	4	4	4	4	**4**	3	3	2	2	2	3.64
30	3	3	3	4	4	3	3	3	3	3	3.43

60	2	2	2	2	2	1	1	1	1	1	2.18
90	3	3	3	2	2	3	3	2	2	2	2.73
120	3	3	3	3	2	2	2	2	1	1	2.86
150	3	2	2	3	3	2	3	1	1	1	2.55
180	3	3	1	3	1	2	1	1	1	1	2.36
210	2	2	2	3	2	2	1	1	1	1	2.29
240	3	3	3	3	3	3	2	3	2	2	3.24
270	3	2	2	3	3	2	2	2	2	1	2.76
300	2	2	2	2	2	2	2	2	2	1	2.36
330	2	2	1	1	1	1	1	1	1	1	1.77
360	2	1	3	2	1	2	1	1	1	1	2.18

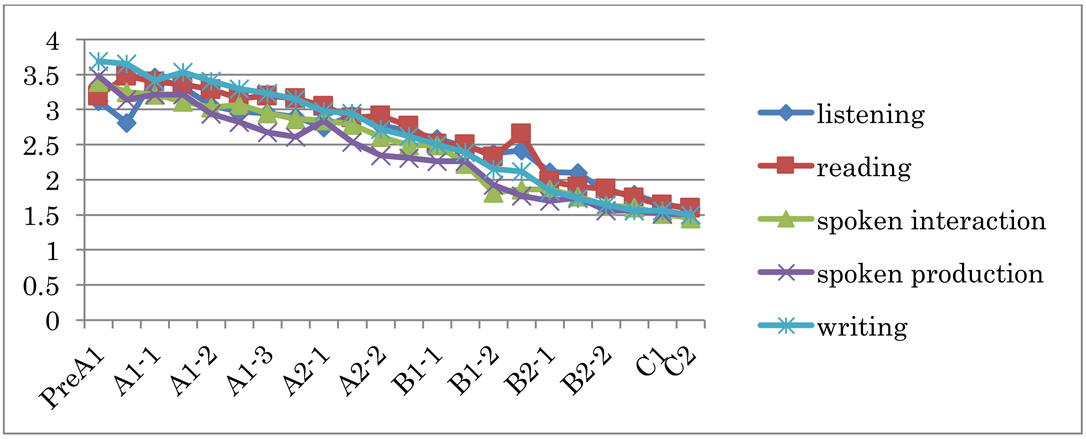
As Table 3 indicates, there is a variation of responses regarding consistency of self-rating levels within individuals and between individuals when EIKEN scores are compared. For example, self-ratings for Can-Do descriptor levels among the respondents are not strongly related to EIKEN test scores. The 1st ranked respondent (highest level sample) chose ‘strongly agree (4)’ for Q17. The average points for the 1st and 30th ranked respondents are higher (respectively ‘3.64’, ‘3.43’) than other respondents. The average points of the 330th and 360th ranked respondents (lowest samples) are ‘1.77’ and ‘2.18’. Those four respondents seem to demonstrate self-ratings which were expected from the EIKEN test score ranking. However, the responses of the 30th ranked respondent are not consistent. ‘4’ was chosen for Qs 1-8, then lower point ‘3’ was chosen for Qs 9-15, again, ‘4’ was chosen for Qs Q16-17. Moreover, the 60th ranked respondent tended to choose “moderately agree (3)” or “moderately disagree (2)” for most of the descriptors and the average point is ‘2.18’, which is somewhat lower than those of other respondents. On the other hand, the 240th ranked respondent with EIKEN low score chose ‘strongly agree (4)’ for Q8 and its average point is ‘3.24’ which is somewhat higher than those of other respondents.

4.2 Reliability and relationship of self-ratings for CEFR-J descriptors

To grasp the overall relationship, the line graph in Figure 1 below was created. In the graph, the horizontal axis shows CEFR-J levels corresponding to questions (Qs) . Figure 1 shows that five skill categories form moderately linear association, descending from higher to lower. That is, as the level of each skill category becomes higher, self-ratings indicate less confident perceptions on Can-Do descriptors. For example, responses for A1.1 center closer to around ‘4’ (strongly agree); responses for B1.1 center around halfway '3' (moderately agree) and '2' (moderately disagree), and responses for C2 is almost halfway between '2' (moderately disagree) and '1' (strongly disagree).

There are slightly extreme average points in the graph below. For example, for listening descriptor Question 2, the average is '2.82' which is lower than those for Questions 3-8. Question 2 descriptor states, "I can recognize the letters of the English alphabet, when they are pronounced." Probably the respondents may have misunderstood that they were asked about knowledge of phonetics. Another extreme average point can be found in reading Q 16, '2.65' which is higher than those of Qs 13, 14 and 15. The descriptor states, "I can understand the plot of longer narratives written in plain English." The respondents may have perceived 'understanding of the narrative (story)' relatively easier, as compared to understanding of texts of internet and reference book (Q15), texts of instruction for games and application (Q14), and texts of newspapers and magazines (Q13). Further in-depth investigation is needed for extreme values in the graph.

Figure 1: Relationship among average self-ratings of five skill categories (N=389)



Next, Cronbach's alpha was utilized to examine internal consistency among the self-ratings of the five skill categories. Cronbach's alpha value among them is '0.872', which shows strong internal reliability among the five skill categories. The result suggests that self-ratings among the five skill categories are statistically reliable, when the average self-ratings of individual respondents for each skill category were compared.

To summarize the results above, when in-depth investigation of the self-ratings of the individuals is conducted, the results indicate that individuals' self-ratings are not consistent. Also, the self-ratings are not strongly related to hierarchy of EIKEN scores. There is a variation of self-ratings for CEFR-J Can-Do descriptors, which is congruent with previous studies by Jones (2002) and Runnels (2014) discussed above. When the individuals in the whole group are statistically analyzed, the relationship between English placement test (EIKEN) scores and the average self-ratings of the CEFR-J Can-Do descriptors is strong.

5. Conclusion and implications

This study reviewed literature related to CEFR, the seemingly most controversial language scale framework in the 21st century, and its Japanese version of the CEFR. This study also conducted the empirical validation of the CEFR-J. This section discusses the results of the study and seeks implications for university English programs in Japan.

When in-depth investigation of the self-rating raw data was conducted, the results show contradictory evidence that there is a variation of self-rating responses within individuals and that individual's responses are not necessarily related to English proficiency test (EIKEN) score. The statistical analysis (Pearson's R) examining the relationship between self-ratings and EIKEN scores also supported the evidence that there is not a strong relationship between the two. However, internal reliability of self-ratings between the five skill categories in this study was found to be strong, using statistical analysis (Cronbach's alpha), when average self-ratings were examined in the group. This result shows that self-ratings of CEFR-J Can-Do descriptors between each skill category are fairly trustworthy.

The researcher makes the following assumptions to interpret this contradictory evidence. CEFR-J Can-Do descriptors may be reliable when they are compared in the group. This implies that language educators may be able to use the CEFR-J Can-Do descriptors effectively to evaluate an entire whole English program regarding the outcome of teaching. On the other hand, individuals show variation in responses of CEFR-J Can-Do descriptors. This may imply that CEFR-J is not reliable measurement method for individual language learning.

Due to time constraints, a variation of responses caused by individual difference was not pursued sufficiently. Further qualitative studies need to be conducted to explicate hidden reasons which cause individual variation in self-ratings.

References

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

ALTE. (2002). *The ALTE CAN DO PROJECT*. Retrieved on March 30, 2015 from (http://www.alte.org/attachments/files/alte_cando.pdf)

Council of Europe. (2001). *The Common European Framework of References for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

ETS. (2014). *Test and score data summary for TEOFL iBT tests (2013, January-2013 December tests)*. Retrieved on March 14, 2015 from (http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf)

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66 (4), 477-485.

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

Jones, N. (2002). Relating the ALTE framework to the Common European Framework of Reference. *THE ALTE CAN DO PROJECT*. Retrieved on March 27, 2015 from (http://www.alte.org/attachments/files/alte_cando.pdf)

Little, D. (2010). The European language portfolio and self-assessment: Using "I can" checklists to plan, monitor and evaluate language learning. In Schmidt, G.S.,

Naganuma, N., O'Dwyer, F., Alexander, E. and Kazuni, S. (Eds.) *Can do statements in language education in Japan and beyond –Applications of the CEFR* (pp.157-166). Tokyo: Asahi Press.

Ministry of Education. (2014). *Eigo kyouiku no arikata ni kansuru yushikisha kaigi saishuhoukoku (English education reform working group final report July, 2014)*. Retrieved on March 26, 2015 from (http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2014/08/20/1351000_01.pdf)

Negishi, M., Takada, T. & Tono, Y. (2012). A progress report on the development of the CEFR-J. *Studies in Language Testing* 36: 137-157.

North, B., Ortega, A., & Sheehan, S. (2010). *A core inventory for general English, British Council/EAQUALS*. Retrieved April 20, 2014 from (http://www.teachingenglish.org.uk/article/british-council-eaquals-core-inventory-general-english-0)

North, B. (2014). *The CEFR in practice*. Cambridge: Cambridge University Press.

Runnels, J. (2013). Examining the difficulty pathways of can-do statements from a localized version of the CEFR. *Journal of Applied Research on the English language*, 2(1), 25-32.

Runnels, J. (2014). An Exploratory Reliability and Content Analysis of the CEFR-Japan's A-Level Can-Do statements. *JALT Journal*, Vol.36, No.1, 69-89.

Tokeshi, M. (2013). TOEFL iBT ni okeru speaking sokutei to writing sokutei no jitsuyousei, shinraisei, datousei [Practicality, reliability and validity of speaking and writing in TOEFL iBT]. *Meio University Bulletin*, No.19, 65-76.

Tono, Y. et al. (2012). CEFR-J Ver.1. Retrieved April 10, 2014 from (http://www.cefr-j. org/download.html.)

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

van Ek, J.A. & Trim, JLM. (2001a). *Waystage.* Cambridge: Cambridge University Press.

van Ek, J.A. & Trim, JLM. (2001b). *The threshold level*. Cambridge: Cambridge University Press.

van Ek, J.A. & Trim, JLM. (2001c). *Vantage*. Cambridge: Cambridge University Press.

Weir, C.J. (2005). Limitations of the Common European Framework for developing comparative examinations and tests. *Language Testing,* 22 (3), 281-300.

Wilkins, D.A. (1976). *National syllabuses*. Oxford: Oxford University Press.