<table>
<thead>
<tr>
<th>著者</th>
<th>Jean-Pierre Joseph Richard</th>
</tr>
</thead>
<tbody>
<tr>
<td>雑誌名</td>
<td>グローバルマネジメント</td>
</tr>
<tr>
<td>巻</td>
<td>21-32</td>
</tr>
<tr>
<td>ページ</td>
<td>2020-08</td>
</tr>
<tr>
<td>発行年</td>
<td><a href="http://doi.org/10.32288/00001307">http://doi.org/10.32288/00001307</a></td>
</tr>
</tbody>
</table>
| URL        | Creative Commons: •
|            | http://creativecommons.org/licenses/by-nc-nd/3.0/deed.ja |
Investigating CEFR-J Self-Assessment and TOEIC Listening and Reading Scores

Jean-Pierre Joseph Richard

Abstract (要約)
The Common European Framework of Reference (CEFR; Council of Europe, 2001) has influenced language education in Japan (Tono, 2017). Negishi, Takada, and Tono (2012) found that approximately 80% of Japanese learners of English are within a narrow basic-user range on the CEFR. Consequently, the CEFR-J was developed for language education in Japan. With English majors, Runnels (2016) observed that hierarchical ordering of the CEFR-J levels was generally as predicted; however, the A- and B-level CEFR-J ratings had weak to near-negligible correlations with TOEIC scores (listening $r = .23$; reading $r = -.14$). This current study attempted to replicate Runnels but with non-English majors. Participants ($N = 54$), first- and second-year students at one public, regional university, completed an online form with the CEFR-J self-assessment descriptors for listening and reading within one week of an end-of-year TOEIC listening and reading test. Compared with Runnels, the predicted hierarchical ordering was more consistent and correlations between A- and B-level CEFR-J ratings and TOEIC scores were stronger (listening $r = .29$; reading $r = .50$). Based on the inconclusive findings compared with Runnels, institutions should proceed with caution if using TOEIC scores and the CEFR-J ratings as language proficiency measures.
Sheehan, 2010). In Japan, for example, the CEFR is used by Nihon Hoso Kyokai for language education and foreign language radio and TV programming (Tono & Negishi, 2012). In addition, owing to the influence of the CEFR, the Japanese Ministry of Education is reforming English language and teaching in Japan based on the CEFR for elementary and secondary school, as well as for entrance examinations (Tono, 2017).

The CEFR “describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively” (Council of Europe, 2001, p. 1). The CEFR uses descriptors, known as can do statements, of increasing difficulty to describe communicative language competencies. These descriptors divide five language skills, (i.e., listening, reading, spoken interaction, spoken production, writing) into three streams of six levels: A1-A2 (Basic User); B1-B2 (Independent User); and C1-C2 (Proficient User) (Council of Europe, 2001). The definitions of proficiency levels found in CEFR allow learners to measure progress across their lifelong learning (Council of Europe, 2001). An example of a CEFR can do statement for listening at the A2 level (Council of Europe, 2001) is: “Can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g., very basic personal and family information...). Can catch the main point in short, clear, simple messages and announcements” (p. 26).

When measuring Japanese learners with the CEFR, Negishi, Takada, and Tono (2012) found that upwards of 80% would be within the CEFR A-level range (i.e., Basic User). Consequently, the CEFR descriptors were modified for English language teaching, learning, and assessment in Japan (Tono, 2014, 2017). To develop the CEFR-J, the first four levels of the CEFR (i.e, A1, A2, B1, and B2) were extended to 10 by adding one Pre-A1 level, and subdividing A into five sub-levels (i.e., A1.1-A1.3, A2.1-A2.2), and B into four (i.e., B1.1-B1.2, B2.1-B2.2) (Tono, 2017). An example of a CEFR-J can do statement for Listening at A1.3 (Tono & Negishi, 2012) is: “Can ask and answer simple questions about very familiar topics (e.g., hobbies, sports, club activities), provided that people speak slowly and clearly with some repetition and rephrasing.”

Runnels (2016) previously investigated the relationship between TOEIC Listening and Reading (TOEIC L&R) test scores and A- and B-level CEFR-J listening and reading self-assessment ratings. Participants in Runnels were English majors studying in a CEFR-informed curriculum. She found that the hierarchical ordering of the CEFR-J levels was generally as predicted; however, adjacent pairs were
frequently misordered. Runnels (2016) also found a weak correlation between TOEIC listening scores and the CEFR-J listening self-assessment ratings ($r = .23$) and a near negligible, albeit negative, correlation for reading ($r = -.14$).

With the exception of research by Runnels, there has been little research which has examined the CEFR-J alongside standardized language tests. Therefore, the focus of the current study is to investigate the relationship between TOEIC L&R test scores and the CEFR-J listening and reading self-assessment ratings. This is important because of the role that TOEIC L&R tests play in Japan. For example, approximately 2.45 million Japanese, including 1.25 million students, completed the TOEIC L&R (Educational Testing Services, 2019) in 2018, and TOEIC L&R scores are frequently used for placement, program evaluation, as well as employment recruitment. With regard to self-assessment ratings, it was noted that the Ministry of Education in Japan is overseeing a reform of English language teaching based on the CEFR. Furthermore, the CEFR-J is used widely in Japan to complement the CEFR (Tono, 2019); and commercial companies, such as Z-Kai, are establishing English-language learning programs based on the CEFR-J (Tono, 2017). This study addresses two research questions. What are the results of non-English majors’ CEFR-J self-assessments for listening and reading? What is the relationship between the learners’ TOEIC L&R scores and the CEFR-J self-assessment ratings for listening and reading?

**Method**

**Instrument**

The CEFR-J currently includes 22 items for each of the five skills of listening, reading, spoken interaction, spoken production, and writing. There are two items for each skill at Pre-A1, all five levels of A, and all four levels of B; and one item each at C1 and C2. For this current study, the 40 CEFR-J items representing A1.1 to C2 for listening and reading were chosen. Pre-A1 items were not chosen because these are intended for young learners or very beginner-students. Each item was scored on a four-point Likert-scale ranging from (1) *I absolutely cannot do this*; (2) *I likely cannot do this*; (3) *I likely can do this*; and (4) *I absolutely can do this*. All items and options were written in Japanese, the participants’ first language.

Two multiple-choice background items, asking major and year at university, and one item for consent to use the data were included. These were the first three items in a Microsoft Form. Participants who did not grant permission to use their data were sent to a “thank you” page with an encouragement to continue to study.
English. Those who granted permission were directed to the 40 CEFR-J items in a sequential order, from A1.1 to C2, with two A1.1 listening items followed by two A1.1 reading items, followed by two A1.2 listening items, and so forth.

**Participants**

Participants were non-English majors at one small, regional public university, *Chubu Public University* (a pseudonym), in central Japan. The participants have four–100-minute required English classes per week in Year 1 and two–to–four–100-minute required classes per week in Year 2, taught by Japanese and non-Japanese tenured and non-tenured faculty members. The CEFR and the CEFR-J frameworks have been part of informal discussions at Chubu Public University; however, neither the CEFR nor the CEFR-J have been formally introduced into the curriculum. Thus, it is likely that most participants were unfamiliar with CEFR-J.

In all, 60 students responded to the questionnaire, of which six opted out of sharing their data. Thus, 54 participants completed the questionnaire and TOEIC L&R. Of these 54, 33 (61.1%) were Year 1 students and 32 (59.3%) were members of the largest department at the university, Economics (a pseudonym). Chubu Public University has approximately equal numbers of first- and second-year students, and the Economics Department comprises approximately 70% of all students; thus, it appeared that Year 1 respondents were over-represented and Economics majors were under-represented in this sample. A chi-squared test of independence (group) of background categorical data that proportionally matches the population of Chubu Public University by year and major would be non-significant: $\chi^2 (1, N = 54) = 0.03$, $p = .86$. However, the chi-squared test of independence (group) of the participants’ background categorical data was highly significant: $\chi^2 (1, N = 54) = 13.87$, $p < .001$, and the phi coefficient ($\phi = .51$) revealed a large effect size; thus, indicating that the participant sample was unevenly distributed across year and major.

TOEIC listening scores ranged from 110 to 465 ($M = 310.19$, $SD = 72.18$), reading scores ranged from 105 to 450 ($M = 253.24$, $SD = 71.90$), and combined scores ranged from 210 to 915 ($M = 563.44$, $SD = 137.88$). Educational Testing Service (n.d.) considers TOEIC listening and reading minimum scores of 110 and 115 to be comparable to CEFR-A2, listening and reading scores of 275 to be comparable to CEFR-B1, and listening and reading scores of 400 and 385 to be comparable to CEFR-B2. Thus, the average participant was B1 for listening and A2 for reading. The distribution of the participants on CEFR are displayed in Table 1.
Table 1. Distribution of Participants’ TOEIC L&R Scores across CEFR Bands

<table>
<thead>
<tr>
<th>TOEIC Test</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>0</td>
<td>19</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>Reading</td>
<td>1</td>
<td>34</td>
<td>15</td>
<td>4</td>
</tr>
</tbody>
</table>

Analyses

The first research question is interested in the CEFR-J self-assessment ratings. To address this first research question, mean ratings of raw scores for CEFR-J-levels were calculated. The second research question is interested in the relationships between TOEIC listening and reading test scores and the CEFR-J self-assessment ratings. To address this second research question, correlations between TOEIC test scores and CEFR-J self-assessment ratings were calculated.

Two of the 40 CEFR-J items, A1.1 listening, “Can understand short, simple instructions... provided they are delivered face-to-face, slowly and clearly” and A1.1 reading, “Can read and understand very short, simple, directions used in everyday life...” had skewness greater than -1.0 (-1.13, -1.02 respectively). A further five items, had moderate skew between -1.0 and -0.5, and between 0.5 and 1.0. The remaining 33 items had skewness values between -0.5 and 0.5, indicating that overall the items were generally fairly symmetrical; and thus, normally distributed. One A1.1-level reading item, “Can understand a fast-food restaurant menu that has pictures...”, had a kurtosis value of -2.02, slightly in excess of what would be expected in a normally distributed sample (>2 or <2). A further five items, had moderate kurtosis values between -2.0 and -1.0, and between 1.0 and 2.0. The remaining 34 items had kurtosis values between -1.0 and 1.0, indicating that the items were appropriately peaked; and thus, normally distributed.

Results

Research Question 1

The mean difficulty ratings for A-, B-, and C-level CEFR-J statements are shown in Table 2. The mean difficulty of the 20 listening items using a four-point Likert scale (4 = ”I absolutely can do this”) was 2.82 (SD = 0.43), and for Reading was 2.77 (SD = 0.42). Removing the C-level items (i.e., most difficult) resulted in mean difficulty scores for listening and reading of 2.94 (SD = 0.44) and 2.90 (SD = 0.44) respectively. As seen in Table 1, for the combined A-, B-, and C-levels, there is a trend from easier to endorse (i.e., A-level, listening: $M = 3.32$, $SD = .43$; reading: $M = 3.29$, $SD = .47$) to more difficult to endorse (i.e., C-level, listening: $M = 1.73$, $SD = .62$; reading: $M$
= 1.66, \(SD = .58\), as predicted by the CEFR-J framework.

Table 2. CEFR-J Mean Difficulty Ratings for Listening and Reading (N = 54).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M)</td>
<td>2.82</td>
<td>2.94</td>
<td>3.32</td>
<td>2.46</td>
<td>1.73</td>
</tr>
<tr>
<td>SE of Mean</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>(SD)</td>
<td>0.43</td>
<td>0.44</td>
<td>0.43</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.09</td>
<td>0.08</td>
<td>0.24</td>
<td>-0.32</td>
<td>0.21</td>
</tr>
<tr>
<td>SE of Skewness</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.79</td>
<td>-0.65</td>
<td>-1.12</td>
<td>-0.41</td>
<td>-0.96</td>
</tr>
<tr>
<td>SE of Kurtosis</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M)</td>
<td>2.77</td>
<td>2.90</td>
<td>3.29</td>
<td>2.42</td>
<td>1.66</td>
</tr>
<tr>
<td>SE of Mean</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>(SD)</td>
<td>0.42</td>
<td>0.44</td>
<td>0.47</td>
<td>0.53</td>
<td>0.58</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.08</td>
<td>0.01</td>
<td>-0.15</td>
<td>0.03</td>
<td>0.40</td>
</tr>
<tr>
<td>SE of Skewness</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.40</td>
<td>-0.21</td>
<td>-0.58</td>
<td>-0.58</td>
<td>-0.47</td>
</tr>
<tr>
<td>SE of Kurtosis</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
</tbody>
</table>

The mean difficulty ratings of the pairs of CEFR listening and reading self-assessment statements for the sub-levels, A1.1-B2.2, and C1-C2, are shown in Table 3. For both listening and reading, there is a trend from easier, A1.1 (listening: \(M = 3.66, SD = .45\); reading: \(M = 3.52, SD = .46\)) to more difficult, C2 (listening: \(M = 1.63, SD = .62\); reading: \(M = 1.66, SD = .63\)). However, mean difficulty ratings for reading at A1.2 and A1.3 were the same (\(M = 3.38\)), and B1.2 mean difficulty reading ratings were 0.01 points higher than B1.1. With the exception of these pairs, all other 18 pairs descend from easiest to most difficult as predicted by the CEFR-J; and thus, the CEFR-J can do self-assessment descriptors performed as intended with these participants.

Table 3. Sub-level CEFR-J Mean Difficulty Ratings for Listening and Reading (N = 54).

<table>
<thead>
<tr>
<th></th>
<th>(M)</th>
<th>SE of Mean</th>
<th>SD</th>
<th>Skew</th>
<th>SE of Skew</th>
<th>Kurt</th>
<th>SE of Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1.1</td>
<td>3.66</td>
<td>0.06</td>
<td>.45</td>
<td>-.84</td>
<td>.33</td>
<td>-.85</td>
<td>.64</td>
</tr>
<tr>
<td>A1.2</td>
<td>3.32</td>
<td>0.07</td>
<td>.53</td>
<td>.10</td>
<td>.33</td>
<td>-.87</td>
<td>.64</td>
</tr>
<tr>
<td>A1.3</td>
<td>3.29</td>
<td>0.07</td>
<td>.51</td>
<td>.18</td>
<td>.33</td>
<td>-.69</td>
<td>.64</td>
</tr>
<tr>
<td>A2.1</td>
<td>3.22</td>
<td>0.08</td>
<td>.58</td>
<td>-.43</td>
<td>.33</td>
<td>.41</td>
<td>.64</td>
</tr>
<tr>
<td>A2.2</td>
<td>3.12</td>
<td>0.07</td>
<td>.52</td>
<td>.11</td>
<td>.33</td>
<td>-.01</td>
<td>.64</td>
</tr>
<tr>
<td>B1.1</td>
<td>2.95</td>
<td>0.08</td>
<td>.58</td>
<td>.11</td>
<td>.33</td>
<td>.10</td>
<td>.64</td>
</tr>
<tr>
<td>B1.2</td>
<td>2.60</td>
<td>0.09</td>
<td>.63</td>
<td>-.10</td>
<td>.33</td>
<td>.67</td>
<td>.64</td>
</tr>
</tbody>
</table>
B2.1  2.37  .10  .72  -.31  .33  -.51  .64
B2.2  1.92  .09  .65  .11  .33  -.78  .64
C1    1.83  .09  .69  .24  .33  -.86  .64
C2    1.63  .09  .62  .46  .33  -.61  .64

Reading
A1.1  3.52  .06  .46  -.23  .33  -1.43  .64
A1.2  3.38  .07  .54  -.43  .33  -.31  .64
A1.3  3.38  .07  .51  -.14  .33  -.75  .64
A2.1  3.15  .09  .66  -.37  .33  -.36  .64
A2.2  3.00  .08  .58  .07  .33  -.02  .64
B1.1  2.81  .08  .61  -.37  .33  .81  .64
B1.2  2.82  .09  .64  -.37  .33  .33  .64
B2.1  2.07  .10  .71  .22  .33  -.67  .64
B2.2  1.96  .08  .62  -.16  .33  -.70  .64
C1    1.70  .08  .60  .23  .33  -.55  .64
C2    1.66  .09  .63  .52  .33  -.59  .64

Note. Mean ratings for underlined items did not follow the expected order.

Research Question 2

The combined TOEIC L&R scores correlated with mean difficulty ratings of the CEFR-J listening \((r = .32, p = .017)\) and reading \((r = .47, p < .001)\). Removing the C-level items (i.e., most difficult) resulted in only marginally improved correlations for listening \((\Delta = .01)\) and reading \((\Delta = .02)\). Slightly smaller correlations were observed between TOEIC listening scores only and mean difficulty ratings of the CEFR-J listening \((r = .28, p = .042)\), with a marginally improved correlation \((\Delta = .01)\) after removing C-level statements. Conversely, slightly larger correlations were observed between TOEIC reading scores and mean difficult ratings of the CEFR-J reading, \((r = .49, p < .001)\), with a marginally improved correlation \((\Delta = .01)\) after removing C-level statements. Correlations for A-, B-, and C-level mean difficulty ratings and TOEIC scores are also found in Table 4. In addition to the paired-skilled correlations (e.g., TOEIC reading and CEFR-J reading), Table 4 also displays the correlations for non-paired skills (e.g., TOEIC reading and CEFR-J listening). As seen, the correlations for listening paired-skills in the second column are lower than the correlations for reading paired-skills and lower than the correlations for non-paired skills. Finally, the average correlations, ranging from .24 to .40, are shown in the last row of Table 4.

All correlations in Table 4 are significant at the .05-level, except for all C-levels, and B-level for listening paired-skills. Correlations were small to moderate based on guidelines from Plonsky and Oswald (2014), where .20 = small, .40 = moderate, and .60 = large for research in second language learning. In all, five correlational analyses
were attempted with each combined set of TOEIC scores and the CEFR-J ratings. Thus, after applying the Bonferroni correction, the \( p \)-value for significance was \( p = .01 \). Consequently, four of the correlations for paired-reading skills, three of the TOEIC listening and the CEFR-J reading correlations, and one of the TOEIC reading and the CEFR-J listening correlations were found to be significantly correlated. These are noted with an “*” in the Table 4.

Table 4. Correlations Between TOEIC and CEFR-J Levels (N = 54)

<table>
<thead>
<tr>
<th></th>
<th>Test Skill</th>
<th>CEFR-J Skill</th>
<th>Listening</th>
<th>Reading</th>
<th>Listening</th>
<th>Reading</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>.28 ( p = .042 )</td>
<td>.49 ( p &lt; .001 )^*</td>
<td>.42 ( p = .001 )^*</td>
<td>.34 ( p = .012 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A+B-level</td>
<td>.29 ( p = .037 )</td>
<td>.50 ( p &lt; .001 )^*</td>
<td>.43 ( p &lt; .001 )^*</td>
<td>.35 ( p = .010 )^*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A-level</td>
<td>.28 ( p = .039 )</td>
<td>.49 ( p &lt; .001 )^*</td>
<td>.45 ( p &lt; .001 )^*</td>
<td>.34 ( p = .011 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B-level</td>
<td>.23 ( p = .090 )</td>
<td>.39 ( p = .004 )^*</td>
<td>.33 ( p = .016 )</td>
<td>.29 ( p = .036 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C-level</td>
<td>.13 ( p = .334 )</td>
<td>.14 ( p = .321 )</td>
<td>.06 ( p = .672 )</td>
<td>.17 ( p = .216 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (SD)</td>
<td>.24 (0.07)</td>
<td>.40 (0.15)</td>
<td>.34 (0.16)</td>
<td>.30 (0.08)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5 displays correlations between TOEIC L&R and individual CEFR-J sub-levels for paired and non-paired skills. As with Table 4, correlations between reading paired-skills were strongest and listening paired-skills were weakest. In all, 24 pairs of correlations were significant at the \( p = .05 \)-level. However, 11 correlational analyses were attempted with each combined set of scores and ratings. Thus, after applying the Bonferroni correction, the \( p \)-value for significance was \( p = .005 \). These are noted with an “**” in the Table 5.

Table 5. Correlations Between TOEIC and CEFR-J Sub-Levels (N = 54).

<table>
<thead>
<tr>
<th></th>
<th>Test Skill</th>
<th>CEFR-J Skill</th>
<th>Listening</th>
<th>Reading</th>
<th>Listening</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1.1</td>
<td>.22 ( p = .118 )</td>
<td>.36 ( p = .008 )</td>
<td>.33 ( p = .016 )</td>
<td>.26 ( p = .055 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A1.2</td>
<td>.27 ( p = .045 )</td>
<td>.43 ( p &lt; .001 )^*</td>
<td>.38 ( p = .005 )</td>
<td>.36 ( p = .007 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A1.3</td>
<td>.43 ( p = .001 )^*</td>
<td>.46 ( p &lt; .001 )^*</td>
<td>.42 ( p = .002 )</td>
<td>.41 ( p = .002 )^*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A2.1</td>
<td>.15 ( p = .268 )</td>
<td>.35 ( p = .009 )</td>
<td>.32 ( p = .018 )</td>
<td>.24 ( p = .083 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A2.2</td>
<td>.11 ( p = .425 )</td>
<td>.49 ( p &lt; .001 )^*</td>
<td>.47 ( p &lt; .001 )^*</td>
<td>.15 ( p = .270 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1.1</td>
<td>.27 ( p = .046 )</td>
<td>.38 ( p = .004 )^*</td>
<td>.39 ( p = .004 )^*</td>
<td>.34 ( p = .013 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1.2</td>
<td>.28 ( p = .043 )</td>
<td>.37 ( p = .006 )</td>
<td>.31 ( p = .023 )</td>
<td>.36 ( p = .008 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B2.1</td>
<td>.15 ( p = .383 )</td>
<td>.34 ( p = .012 )</td>
<td>.25 ( p = .075 )</td>
<td>.19 ( p = .162 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B2.2</td>
<td>.11 ( p = .441 )</td>
<td>.17 ( p = .223 )</td>
<td>.13 ( p = .350 )</td>
<td>.10 ( p = .455 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C1</td>
<td>.13 ( p = .339 )</td>
<td>.27 ( p = .050 )</td>
<td>.18 ( p = .184 )</td>
<td>.16 ( p = .246 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>.12 ( p = .391 )</td>
<td>.00 ( p = .983 )</td>
<td>-.07 ( p = .629 )</td>
<td>.16 ( p = .242 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (SD)</td>
<td>.20 (.10)</td>
<td>.33 (.14)</td>
<td>.28 (.15)</td>
<td>.25 (.11)</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

Similar to Runnels (2016), CEFR-J level hierarchical ordering was as predicted, with A-levels being easier to endorse than B-levels, which were easier to endorse than C-levels. In all, two out of 20 adjacent pairs were misordered, fewer than were observed in Runnels despite there being fewer adjacent pairs in Runnels \((k = 16)\). With regard to relationships between TOEIC scores and the CEFR-J ratings, Runnels found mostly negligible-to-small correlations for listening and none-to-negligible for reading, and the reading correlations in Runnels were negative. In this current study, however, TOEIC reading test scores and mean difficulty ratings for the CEFR-J reading statements were more strongly correlated than for listening, and more compared with either of the correlations for non-paired skills. In addition, all correlations were positive, and ranged from negligible to moderate. Ross (1998), as Runnels noted, in his meta-analysis of self-assessment research in second-language and foreign-language acquisition, observed that reading test scores and self-assessments are more likely to correlate with each other, and for these correlations to “appear robust” (Ross, 1998, p. 5), compared with other skills. This is likely because “reading tends to be the skill that is first taught in the foreign language context, ...[and university participants] were most likely very experienced in using their reading skills” (Ross, p. 6). The finding of stronger reading correlations was also observed in the current study, but not in Runnels.

It was also observed that the listening paired-skills’ correlations were smaller than the non-paired skills correlations. It is unclear why this might be so. However, one possibility is the mismatch between listening skills and test formats. The listening section of the TOEIC test requires reading skills to answer the paper-based formatted test; and the CEFR-J listening descriptors are also paper-based, although these descriptors are written in the participants’ first language. Thus, neither the listening test nor the listening self-assessment descriptors match the skill and test format. In contrast, both the reading section of the TOEIC test and the CEFR-J reading descriptors require reading skills to respond; and thus, match the skill and test formats. With regard to each non-paired skills, either the reading section of the TOEIC test or CEFR-J reading descriptors match the skill and test formats. Runnels (2016) also noted “self-assessment is likely to be affected by task difficulty” (p.122) in that easier tasks are more accurately rated than difficult ones. In this current study, for each series of correlations, A-level correlations were strongest, followed by B-level, then C-level. This is, as Runnels highlighted, easier tasks are more accurately rated. It should be noted, however, that the participants in this
current study had higher TOEIC scores than those in Runnels, and correlations were generally higher. It might be that learners in the present study have had more experiences at attempting more difficult tasks as defined by the CEFR-J descriptors, especially for A- and B-levels, and for reading in particular than the participants in Runnels. In this way, previous experiences at more difficult tasks would likely be an important factor in the size of the correlation between the TOEIC test and the CEFR-J descriptors.

Conclusion

The current study reported on an investigation of Japanese learners’ of English self-assessment ratings based on the CEFR-J can do descriptors for listening and reading and TOEIC L&R test scores. The results described here were also briefly compared with the results found in Runnels (2016). In this current study, as with Runnels, the CEFR-J levels performed as predicted, that is, easier items had higher mean difficulty ratings, and more difficult items had lower mean difficulty ratings. However, different from Runnels, the correlations in this current study were strongest for reading and weaker for listening; whereas she found the opposite.

Several limitations were observed regarding the results in this current study when compared with the results in Runnels (2016). First, N-sizes were relatively small in both the former (N = 54) and the latter (N = 57). These smaller sample sizes might contribute to different results observed in both studies. Second, participants in the current study were different from those in Runnels, with non-English majors in a curriculum not informed by the CEFR in the former, and English majors in a curriculum informed by the CEFR in the latter. Observed differences in the results between the two studies might be due to differences between samples, as English majors and non-English majors might perceive the CEFR-J self-assessment can do statements differently. In addition, students studying in a CEFR-informed curriculum might have a greater awareness and understanding of the nuances in the CEFR-J descriptors compared with students studying in a curriculum not informed by the CEFR. However, this does not explain why higher correlations were observed in the current study. Perhaps the participants in Runnels proceeded too cautiously with the CEFR-J descriptors. Third, participants in this current study had larger TOEIC L&R scores. These higher TOEIC test scores likely relate to more experiences with more difficult tasks. If less able students have had fewer experiences with more difficult tasks, then it is likely that they can imagine neither the complexity involved in task mastery nor successful completion of a difficult task. Based on the inconclusive
findings in this paper, especially when compared with Runnels, and based on the limitations noted, Japan-based educational institutions, language educators, and researchers likely need to proceed with caution if using the TOEIC test scores and the CEFR-J self-assessment data as measures of language proficiency.

References (参考文献)


Educational Testing Services (n.d.). TOEIC® program 各テストスコアとCEFRとの対照 [tōikku® puroguramu kaku tesuto sukoa to CEFR to no taishō-hyo, TOEIC® Program test score and CEFR comparison table]. https://www.iibc-global.org/toeic/official_data/toeic_cefr.html


Tono, Y. (2019). Coming full circle—From CEFR to CEFR-J and back. CEFR Journal,