The CEFR-J Hierarchy and its Relationship with TOEIC Listening and Reading

Jean-Pierre Joseph Richard
The University of Nagano

Reference Data

Richard, J. - P. J. (2021). The CEFR-J hierarchy and its relationship with TOEIC Listening and Reading. In P. Clements, R. Derrah, & P. Ferguson (Eds.), *Communities of teachers & learners*. JALT. https//doi.org/10.37546/JALTPCP2020-50

Runnels (2016) investigated the CEFR-J self-assessment sublevel hierarchy and its relationship with TOEIC Listening and Reading (TOEIC L&R). Runnels (a) found learners did not distinguish between adjacent sublevels, and (b) observed mostly small-to-negligible Pearson's *r* correlations between CEFR-J and TOEIC L&R, with slightly stronger correlations for listening. In the current study, participants (*N* = 53) completed a questionnaire (*k* = 36) with statements representing CEFR-J sublevels A1.1 to B2.2. Groups (*i.e.*, A>B) and levels (i.e., A1>A2≧B1>B2) performed as predicted; however, participants mostly did not differentiate between adjacent sublevel items. Small-to-moderate Kendall's τ correlations between CEFR-J and TOEIC L&R were observed, with stronger correlations for reading. Despite the lack of clarity regarding sublevels, learners might interpret the levels as coherent sets (Negishi, 2020). More experiences with reading likely resulted in more robust reading correlations (Ross, 1998). One contribution of this paper is its partial replication of Runnels, with potentially improved methodological tools.

Runnels（2016）は、学習者が隣接するCEFR-Jサブレベルを区別しなかったことを発見し、CEFR-JとTOEIC　L&Rの間のピアソンの相関係数がほとんど無視できる程度であり、リスニングではわずかに強くなることを見いだした。本論において、参加者（N = 53）はCEFR-Jサブレベル（A1.1～B2.2）を表すアンケート（k = 36）に回答した。CEFR-Jのグループ（AとB）とレベル（A1, A2, B1, B2）は予測どおりに実行されたが、参加者は隣接するサブレベルをほとんど区別しなかった。CEFR-JとTOEIC L&Rの間には、小から中程度のケンドールのτ相関係数が見られ、リーディングではやや強い相関が見られた。サブレベルに関する明確さの欠如にもかかわらず、学習者はレベルを一貫したセットとして解釈する可能性があり（Negishi,2020）、読書の経験が多いほど、相関がより強固になると考えられる（Ross,1998）。本論の学術的な貢献は、、Runnelsの研究の一部を方法論的に改善した形で再現したところにある。

Following its publication in 2001 by the Council of Europe, the Common European Framework of Reference (CEFR) has had a large impact on language education globally (Byram & Parmenter, 2012; Piccardo, 2019; Read, 2019). In Japan, for example, the Ministry of Education (MEXT) uses CEFR for informing English-language education (Tono, 2017), and for aligning external language tests (MEXT, 2018). However, most Japanese English learners were *Basic Users* (Negishi, 2012), including 98% of Year 3 high school students at national and public high schools (MEXT, 2015). Thus, for CEFR to be meaningful for English-language education in Japan, it needed to be adapted to the local context. Japanese researchers developed a Japanese version of CEFR, known locally as CEFR-J (Negishi, 2012; Tono, 2013). They added a Pre-A1 sublevel, subdivided A1-A2 into five sublevels and B1-B2 into four sublevels (Negishi, et al, 2013; Tono, 2013). These researchers believed that CEFR-J would allow for fine-tuning of instruction and assessment, and Japanese learners would be able to see progress more clearly (Tono, 2017). See Table 1 for information about CEFR groups and levels, CEFR-J sublevels, and approximate comparisons with selected external tests.

Tokeshi and Gao (2015) attempted to validate the CEFR-J hierarchical levels by investigating the correlations with the Eiken test in Practical English Proficiency (Eiken) as well as the CEFR-J internal relationships (i.e. Cronbach's α). They observed small correlations between CEFR-J and EIKEN scores (e.g., listening, $r = .27$; reading, $r = .29$), but found that the internal relationship (Cronbach's $\alpha = .87$) among the self-assessment ratings was acceptable. Runnels (2016) researched the CEFR-J hierarchy and its correlations with scores from the TOEIC Listening and Reading (TOEIC L&R). She found that the mean score for CEFR-J A-level items (i.e., A1.1-to-A.2.2) was higher than the mean score for B-level items (i.e., B1.1-B2.2); however, few adjacent sublevels were significantly different from each other. She also observed that correlations between TOEIC L&R and CEFR-J sublevels were small for listening ($r = .23$), and negligible and negative for reading ($r = -.14$).

Table 1
CEFR and CEFR-J Levels and Score Comparisons with Selected Tests

Group: User	Level (Name)	CEFR-J sublevel [a]	Eiken [b]	IELTS	TOEIC L&R + S&W
C: Proficient	C2 (Mastery)	(C2)		8.5 - 9.0	
	C1 (Advanced)	(C1)	Grade 1	7.0 - 8.0	1845 - 1990
B: Independent	B2 (Vantage)	B2.2	Grade Pre-1	5.5 - 6.5	1560 - 1840
		B2.1			
	B1 (Threshold)	B1.2	Grade 2	4.0 - 5.0	1150 - 1555
		B1.1			
A: Basic	A2 (Waystage)	A2.2	Grade Pre-2		625 - 1145
		A2.1			
	A1 (Breakthrough)	A1.3	Grade 3		320 - 620
		A1.2			
		A1.1			
		PreA1			

Note. Score comparisons are from MEXT (2018, March).

a = CEFR C1 and C2 levels were left unchanged in CEFR-J.

b = MEXT (2014) benchmarks for Japanese English teachers (Grade Pre-1), Japanese high school graduates (Grade 2), and junior high school graduates (Grade 3).

However, Tokeshi and Gao (2015), did not clarify how the reliability analysis was completed, yet it appears that all five skills were combined in one analysis. This is problematic because Cronbach's α should not be interpreted as a measure of internal consistency when different subskills are combined (Field, 2020). In addition, Tokeshi and Gao (2015) provided no details of the participants' performance on Eiken, the language test that was used in their correlational analyses. In Runnels (2016), B-level items were frequently unreported in favor of A-level items, and a one-way ANOVA was used instead of a repeated measures ANOVA. Moreover, it is possible, as seen in the current study, CEFR-J data distributions in the above studies were nonnormal, requiring nonparametric tests. This might be a feature of the hierarchical nature of CEFR-J can-do descriptors; however, in general, neither paper reported descriptive statistics for CEFR-J items, including no mention of normality of data.

With the exception of the above two studies, there is limited research attempting to validate CEFR-J. This is pertinent because of the expanding number of curricula informed by CEFR and CEFR-J and the role that external language tests, in particular TOEIC L&R, play, including in admissions, program evaluation, hiring and promotion (Im et al, 2019). Based on Runnels (2016), there are two principal research questions in this current study: (1) Are adjacent CEFR-J levels hierarchically ordered and significantly different? and (2) What are the relationships between CEFR-J listening and reading self-assessment ratings and TOEIC L&R test scores? Although these two research questions are based on Runnels, this paper should be seen as a partial replication of Runnels. Partial in that CEFR-J can-do self-assessment statements are used, as is the TOEIC L&R; however, the background of participants and the type of statistical analyses used are different. Nonetheless, this paper should still be viewed as a replication study, and these are important as they "enhance the reliability of findings in the field and strengthen theoretical claims" (p. 166). Also note that an earlier version of this paper was published in The Global Management of Nagano (Richard, 2020).

Methodology

Context and Participants

This current study was localized at one public university, Shinano University (a pseudonym), in central Japan, where first-year students have four 100-minute English classes per week, second-year students have two-to-four, and all second-year students participate in a short-term study abroad program. The impetus for this study was a desire by the author to introduce CEFR and CEFR-J for future alignment of the curriculum. Institutional clearance was granted for this study.

Within one week of a year-end TOEIC L&R test, first- and second-year students were emailed Japanese explanations of this study, including informed consent details, and a link to a form for participating. Only if students gave consent were data collected. The form included L1-written CEFR-J items (k = 36) representing A1.1 to B2.2 for listening (k = 18) and reading (k = 18). These were scored on a four-point Likert-scale, translated from Japanese as (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*. In this current study the term *group* refers to A or B supralevels and *level* refers to A1, A2, B1, or B2 found in the original CEFR, and *sublevel* refers to individual CEFR-J items from A1.1 to B2.2.

The participants (N = 53), non-English majors, represented 12% of students in first- (n = 32) and second- (n = 21) year. Their mean TOEIC listening scores (M = 314) were higher than their mean TOEIC reading scores (M = 256). Educational Testing Service (n.d.) provides a conversion chart between TOEIC L&R scores and CEFR levels. Based on this, Table 2 summarizes the participants' scores and comparable CEFR levels. The mode was CEFR B1 for listening with a subgroup at A2, and CEFR A2 for reading with a subgroup at B1. The reading scores had three high-scoring outliers, none of whom had ever been overseas except as part of the university's study abroad program. These three were deemed to be from the sample population.[1] In addition, the reading test scores were moderately skewed. Tabachnick and Fidell (2007) recommended a square root transformation for moderately positively skewed variables. Following transformation, skewness improved for the reading test scores. See Appendix A for descriptives for TOEIC Listening and Reading.

Table 2

CEFR Levels (N = 53) from Converted TOEIC L&R Bands

CEFR Comparison	TOEIC Bands: Listening	Listening (n)	TOEIC Bands: Reading	Reading (n)
A1	60-105	0	60-110	0
A2	110-270	18	115-270	34
B1	275-395	29	275-380	15
B2	400-485	6	385-450	4
C1	490-	0	455-	0

Analyses

Appendix B displays the raw descriptives for listening and reading mean scores for CEFR-J item groups and levels, and Appendix C displays these descriptives for listening and reading for CEFR-J sublevels. Z-scores for skewness and kurtosis were mostly within acceptable range (e.g., Z-skew < 1.96); however, most variables for CEFR-J levels and sublevels had significant Shapiro-Wilk's values, indicating that distributions were significantly different from normal. In addition, distribution plots and Q-Q plots clearly indicated nonnormal distributions for most of CEFR-J variables. Data transformations were applied following Tabachnick and Fidell (2007). For most CEFR-J variables, the Shapiro-Wilk's p-value and plots did not improve and worsened in some cases. Thus, nonparametric tests were used for most of the subsequent analyses, with all analyses conducted using JASP, Version 0.13.1 (JASP Team, 2020).

For research question 1, three sets of tests were completed for both listening and reading, one for groups, levels, and sublevels. The A and B groups for listening were parametric (Shapiro-Wilk W = .983, p = .643); for reading they were nonparametric (Shapiro-Wilk W = .948, p = .023). The two listening groups (i.e., A and B) were compared using a paired-sample Student's t-test, whereas the reading groups (i.e., A and B) were also compared using a paired-sample Wilcoxon t-test. The levels (e.g., A1 vs. A2) and sublevels (e.g., A1.1 vs. A1.2) were analyzed using a Friedman one-way repeated measure analysis of variance by ranks (i.e., Friedman's Test), a nonparametric repeated measures ANOVA. This omnibus test is used to test whether there are differences in means between three or more groups, with the same participants in each group measured under different conditions (Lund & Lund, 2018). In this study, the different conditions are represented by the hierarchical CEFR-J levels, and our interest is in differences between adjacent levels or sublevels. Effect sizes for the nonparametric RM-ANOVAs are reported in Kendall's W, which ranges from 0 to 1, with higher values indicating larger effects. For identifying differences in paired-comparisons, Conover's post hoc test was calculated with the Holm-Bonferroni test applied to control for the family-wise error rate (Goss-Sampson, 2020). Robust statistics are unavailable for Conover test results in JASP; thus, paired sample Wilcoxon t-tests for data were also run to provide confidence intervals for effect sizes.

For research question 2, Kendall's tau (τ), a nonparametric correlation, applied when many values have the same score as was the case with CEFR-J data, was used. Importantly, τ is thought to be a better estimation of the correlation in the sample population (Howell, 1997); however, the correlation coefficient for τ is typically smaller[2] than that for r. Gilpin (1993) provided a useful conversion table from τ to r.

Results

RQ1. Are adjacent CEFR-J levels hierarchically ordered and significantly different?

The mean score of the A-group items for listening was significantly higher than the B-group, with a very large effect (d = 2.018) and wide confidence intervals for the parametric Student's t-test. For reading, the effect size for the nonparametric Wilcoxon test was also large.

Table 4
Paired Samples T-Tests for CEFR-J Groups (N = 53)

					95% CI for Effect Sizes	
Skill	Test	Statistic (df)	p	Effect size	Lower	Upper
Listening	Student	14.690 (52)	< .001	2.018	1.543	2.486
Reading	Student	13.086 (52)	< .001	1.797	1.357	2.231
Reading	Wilcoxon	1378.000	< .001	1.000	1.000	1.000

Note. For the Student's t-test, effect size is given by Cohen's *d*. For the Wilcoxon test, effect size is given by the matched rank biserial correlation.

The mean scores for the four levels (e.g., A1 vs. A2) for listening and reading were compared. For listening, the nonparametric RM-ANOVA was significant, X^2 (3) = 128.800, *p*. <.001, Kendall's *W* = .707. Conover test pairwise comparisons indicated that all four adjacent CEFR-J levels were significantly different at each level. Wilcoxon test effect sizes for pairwise comparisons were large with narrow confidence intervals (CIs) - see Appendix D. For reading, the nonparametric RM-ANOVA was significant, X^2 (3) = 125.582, *p*. <.001, Kendall's *W* = .650. Conover test pairwise comparisons indicated that three of the four adjacent CEFR-J levels were significantly different from each other. Wilcoxon test effect sizes for pairwise comparisons, for A1-A2 and B1-B2, were large with narrow CIs. Note that the effect size for the nonsignificant A2-B1 pairwise comparison was medium and its CIs did not cross zero.

The mean scores for the nine sublevels (e.g. A1.1 vs. A1.2) for listening and reading were compared. For listening, the nonparametric RM-ANOVA was significant, X^2 (8) = 276.960, *p*. <.001, Kendall's *W* = .576. Conover test pairwise comparisons indicated that three of the eight adjacent sublevel pairs were significantly different from each other. Wilcoxon test effect sizes for these three pairwise comparisons, A1.1-A1.2, B1.1-B1.2, and B2.1-B2.2, were large with narrow CIs. Two other pairwise comparison also had effect sizes whose CIs did not cross zero (i.e., A2.2-B1.1, B1.2-B2.1) - see Appendix E. For reading, the nonparametric RM-ANOVA was significant, X^2 (8) = 282.790, *p*. <.001, Kendall's *W* = .515. Conover test pairwise comparisons indicated that one of the eight pairs of CEFR-J sublevels was significantly different from the other. Wilcoxon test effect size for the B1.2-B2.1 pairwise comparison was large with narrow CIs. In addition, four other comparisons had effect sizes whose 95% did not cross zero (i.e., A1.1-A1.2, A1.3-A2.1, A2.1-A2.2, A2.2-B1.1).

RQ2. What are the relationships between CEFR-J listening and reading self-assessment ratings and TOEIC L&R test scores?

Appendix F displays the Kendall's τ coefficients between the means of all CEFR-J items per skill and TOEIC scores, as well as the means for CEFR-J groups, levels, and sublevels. For all, the τ coefficient for listening (τ = .197, *p* = .042) was smaller than for reading (τ = .322, *p* < .001). Note also that the CIs for the correlation for listening crosses zero. For groups, levels, and sublevels, reading variables had stronger correlations, than listening variables, with the exception of sublevel A1.3. Of the 16 variables for each paired skill, only five for listening, including two at the sublevel, were significant compared with 12 for reading, including six at the sublevel. Importantly, 10 of the 16 variables for listening had τ correlation coefficient CIs which crossed zero, compared with only two for reading, once again highlighting the weaker correlations for listening. Note also that no correlations were significant beyond B1.2, the upper band for participants in this study based on TOEIC L&R scores.

Discussion and Conclusion

This study was a partial replication of Runnels (2016), with differences between the two studies in population (e.g., major, curriculum) and analyses (e.g., RM-ANOVA instead of a one-way ANOVA). The hierarchical ordering of CEFR-J groups, levels, and sublevels for listening and reading were investigated, as were the relationships between CEFR-J and TOEIC scores. Most data variables were nonnormal, and consequently most analyses used nonparametric tests.

For research question 1, similar to Runnels (2016), for both skills, the mean of the A-group items was significantly higher than that of the B-group items. A second data run-investigating the differences in the mean of the items at each adjacent level (e.g., A1-A2) revealed all pairs for listening and two of three pairs for reading were hierarchically ordered and significantly different. Importantly, effect sizes were medium-to-large, and the CIs did not cross zero. Note also that Runnels did not report the results at this level. Finally, at CEFR-J sublevels three adjacent pairs for listening and one for reading were significantly different from each other, although five pairs each for listening and reading had effect sizes which did not cross zero.

Thus, similar to Runnel's (2016) conclusion, CEFR-J groups functioned as intended; that is, A-group items were easier to endorse than B-group items. Moreover, this held true for the levels (i.e., listening: A1 > A2 > B1 > B2; reading: A1 > A2 ≧ B1 > B2). However, the participants in this current study were mostly unable to distinguish

between the proposed difficulty levels in adjacent CEFR-J sublevel items. As noted, the participants in this current study differed from those in Runnels in important ways. The participants in Runnels were English majors in a CEFR-informed curriculum, with training in the development of their own can-do statements and in self-assessment. The participants in the current study were not English majors, were not in a CEFR-informed curriculum, and had no training in developing can-do statements nor in self-assessment. Despite these differences, similar results with Runnels, for the first research question, were obtained.

These similar results, that is observed differences for groups and levels but not sublevels, despite different contexts might reflect the CEFR-J hierarchy itself. Negishi (2020, and personal communication) indicated that it might be difficult for Japanese learners to distinguish between adjacent sublevels, yet he argued that these sublevels remain informative as targets for individual learners. Moreover, according to Negishi, the levels (e.g., A2, B1) are likely interpreted by learners as being near coherent sets of items, and that learners sense the quality or state of each set (e.g., the *A1-ness* of A1-level items), meaning that learners view the items at a particular level (e.g., A1) as different from those at an adjacent level (e.g., A2). The results from the current study, that is the hierarchical ordering of adjacent levels (i.e., for A1 > A2 ≧ B1 > B2), might support this argument.

Regarding research question 2, on first appearance, the correlation observed in this study for listening (τ = .20) was near comparable with the correlations that were observed by Runnels (2016) and Tokeshi and Gao (2015). Likewise, it seems that the correlation in this study for reading (τ = .32) was similar to that as observed by Tokeshi and Gao. However, the correlations in this current study were Kendall's τ, not Pearson's *r* as reported by these previous authors. Following Gilpin (1993), the converted τ coefficients from this current study to *r* coefficients would be small for listening (*r* = .31) and medium-sized (*r* = .50) for reading, while those from Runnels and Tokeshi and Gao were small for both skills[3]. Second, Eiken was used by Tokeshi and Gao, whereas TOEIC L&R was used by Runnels and in the current study. These differences need to be considered when comparing results.

Considering the A1-B2 levels, the strongest correlation for the paired skill of CEFR-J listening and TOEIC Listening was at B1; for CEFR-J reading and TOEIC Reading it was at A2 and B1. Based on TOEIC score bands (Table 2), the average (i.e., mode) participant in this current study was at B1 for listening and A2 for reading. Thus, comparable CEFR bands based on TOEIC scores for listening and reading aligned with the strongest correlations between the paired skills, implying a certain level of concurrent validity. At the sublevel, stronger correlations were almost exclusively observed between the paired skill of reading rather than for listening. The correlational findings from the current study were congruent with those reported in the meta-analysis by Ross (1998), that is, reading self-assessment scores were relatively more accurate than listening self-assessment; and conversely, they are incongruent with those observed by Runnels (2016) in which reading correlations were small to negligible and negative.

The differences in ability of the participants in Runnels (2016) compared with those in this current study, as measured by TOEIC L&R, might help to account for some of the discrepancies in the correlation results. Higher TOEIC performance likely indicates that many participants in this current study have had more experiences engaging with more difficult language learning tasks. Learners with more experiences in successful task engagement likely have a greater awareness of the meaning of a particular CEFR-J can-do statement at a particular level. As Ross (1998) noted, "the degree of experience learners bring to the self-assessment context influences the accuracy of the product" (p. 16). Thus, the more you know, the more you know you know, and the opposite is also true, suggesting that CEFR-J can-do descriptors are most accurate when directly targeted at particular learners with particular experiences and proficiencies. Further evidence of this was observed where no correlations beyond the B1-level, the upper band for most participants in this current study, were significant.

Immediate comparisons between the results in this current paper and those in Runnels (2016) and Tokeshi and Gao (2015) have been discussed; however, it is important to note that based on observations from this current study, it is quite possible that data in these previous papers were nonnormal and data transformations or nonparametric tests were necessary. Therefore, comparisons need to be treated with caution and viewed skeptically. Furthermore, one important limitation with the current study is the sample size (*N* = 53), and in particular, its range, which might have stunted the results. As most participants in the current study are at the B1-level for listening and A2-level for reading as measured by TOEIC L&R, differences between adjacent CEFR-J sublevels might not be readily apparent. A wider range of abilities might have resulted in greater separation of adjacent pairs and stronger correlations. Runnels (2016) and this current study were small in scale and localized at individual regional universities. Thus, more research, with larger and more varied samples, is needed. One further limitation of this current study, and of Runnels, is that CEFR-J sublevel scores were reported as means for pairs of items. The averaging of scores for item pairs is problematic and assumes that participants view each item for each pair as approximately equal. As the performance of individual items remains unknown, future analysis, including Rasch analysis, should investigate the individual items.

Returning to external testing, a potential mismatch between curriculum and such testing exists at many educational institutions. Moreover, many Japanese learners might wish for a high score on external tests but studying English generally, and for external tests in particular, might be what Richard and Uehara (2017) described as idealized effort, a wished-for intention that remains unacted upon. In contrast to this potential mismatch, CEFR and CEFR-J informed curricula are action oriented with learners as social agents (Council of Europe, 2018). Skills are developed with learners as active participants in a purposeful learning process. English programs, including at Shinano University, might benefit by aligning curricula with CEFR and CEFR-J; however, internal (e.g., program evaluation) and external (e.g., job-hunting) testing requirements remain. The weak-to-moderate correlations observed in this study between CEFR-J self-assessment scores and scores from the TOEIC L&R might be perceived unfavorably by certain stakeholders. Importantly, CEFR "exists primarily to help language professionals and language learners achieve their goals more successfully, to help us think about how and what we teach and learn" (Frost & O'Donnell, 2015, p. 4), and CEFR-J sublevels are also intended to benefit learning and teaching (Tono, 2017). Thus, CEFR and CEFR-J are not designed to be primarily assessment tools, but rather assessment should be from a "complex and dynamic perspective, in a constant interdependent relationship with teaching and learning" (Piccardo, 2019, p.2). Importantly, this study is likely limited by the range of its sample, yet despite this limitation, CEFR-J and the TOEIC L&R were found to have a certain level of concurrent validity. One implication of this is that CEFR (or CEFR-J) informed curriculum might still achieve hard or soft targets with regard to the TOEIC L&R.

Finally, Larson-Hall (2016) has indicated that replication studies enhance reliability in this field. This paper, as a partial replication of Runnels (2016), is one small step to enhancing reliability. Moreover, this paper provided a rich description of its methodology, including its descriptives with robust statistics, and its analyses, as called for by both Plonsky (2015) and Larson-Hall in order for our field to move from old statistics to new statistics. Future quantitative research papers in our field should apply similar methodologies.

Notes

1. As indicated, all second-year students at Shinano University take part in a study abroad program; thus, these three outliers were similar in this regard as their cohort mates (n = 21) in this study.
2. Kendall's τ results in a coefficient that is approximately 66-75% the size of Pearson's r (Strahan, 1982); approaching a ratio 3:2 (Fredericks & Nelson, 2006). The formula for converting τ to r is: $r = .5(\pi\tau)$ (Walker, 2003).
3. Educational Testing Services (2019) reported correlations as large as r = .57 and r = .52 between a self-assessment questionnaire composed of items related to practical English language tasks and TOEIC Listening and TOEIC Reading, respectively.

Bio Data

Jean-Pierre Joseph Richard is a member of the Faculty of Global Management at the University of Nagano. His research interests include measurement and methodology. <richard.jean-pierre@u-nagano.ac.jp>

References

Byram. M., & Parmenter, L. (Eds.) (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Multilingual Matters.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg, France: Council of Europe. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Educational Testing Services (n.d.). TOEIC® Program各テストスコアとCEFRとの対照表 [TOEIC® Program test score and CEFR comparison table]. https://www.iibc-global.org/toeic/official_data/toeic_cefr.html

Educational Testing Services. (2019). *Score User Guide: TOEIC Listening & Reading Test*. https://www.ets.org/s/toeic/pdf/toeic-listening-reading-test-user-guide.pdf

Field, A. (2020). *Discovering Statistics Using SPSS* (5th ed.). Sage Publications.

Frost, D., & O'Donnell. (2015). Réussite : Être ou ne pas être B2 : Telle est la question. (Le projet ELLO Étude longitudinale sur la langue orale) [Success B2 or not B2 – That is the question. (The Étude longitudinale sur la langue orale (ELLO) project)]. Cahiers de l'Apliut, XXXIV(2). https://journals.openedition.org/apliut/5195

Fredericks, G. A., & Nelson, R. B. (2006). On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *Journal of Statistical Planning and Inference, 137,* 2143–2150. https://doi.org/10.1016/j.jspi.2006.06.045

Gilpin, A. R. (1993). Table of conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement, 53*, 87–92. https://doi.org/10.1177/0013164493053001007

Goss-Sampson, M. A. (2020, May). *Statistical analysis in JASP: A guide for students. JASP v0.12*. https://doi.org/0.6084/m9.figshare.9980744

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Duxbury Press.

Im, G. - H., Shin, D., Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9,* 14. https://doi.org/10.1186/s40468-019-0089-4

JASP Team. (2020). *JASP* (Version 0.13) [Computer software]. https://jasp-stats.org/

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lund, A., & Lund, M. (2018). Repeated Measures ANOVA. *Lærd Statistics*. https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

MEXT. (2014).「グローバル化に対応した英語教育改革実施計画」について [About "English Education Reform Implementation Plan for Globalization"]. https://www.mext.go.jp/a_menu/kokusai/gaikokugo/1343704.htm

MEXT. (2015). 平成26年度 英語力調査結果(高校3年生)の概要(詳細版) [Summary of 2014 English Proficiency Survey Results (3rd year high school students) (Detailed version)]. https://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2015/07/03/1358071_02.pdf

MEXT. (2018, March). 各資格・検定試験とCEFRとの対照表 [Comparison Table Between Levels/Tests and CEFR]. https://www.mext.go.jp/b_menu/houdou/30/03/__icsFiles/afieldfile/2019/01/15/1402610_1.pdf

Negishi, M. (2012, March). *The development of the CEFR-J: Where we are, where we are going*. http://www.tufs.ac.jp/common/fs/ilr/EU_kaken/_userdata/negishi2.pdf

Negishi, M. (2020, October 23-25). *What's done and what's not done? The use of the CEFR in Japan. The praxis of teaching, learning, and assessment with CEFR and CLIL* (Online). https://cefrjapan.net/events

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE [Association of Language Testers in Europe] Krakow Conference, July 2011* (pp. 135–63). Cambridge University Press. https://www.alte.org/resources/Documents/SILT%2036%20FOR%20PRINT.pdf

Piccardo, E. (2019). *TIRF language education in review – The Common European Framework of Reference (CEFR) in language education: Past, present, and future.* TIRF & Laureate International Universities.

Plonsky, L. (2015). *Advancing Quantitative Methods in Second Language Research.* Routledge.

Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific Region. *Language Education and Acquisition Research Network Journal, 12*(1), 12–18.

Richard, J.-P. J. (2020). CEFR-J can do self-assessment and TOEIC Listening and Reading scores. *The Global Management of Nagano, 3,* 21-32.

Richard, J.-P. J., & Uehara, S. (2017). Using content analysis to identify L2 motivation and efforts to learn English. *Tokyo JALT Journal, 4,* 29-37.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15,* 1–20. https://doi.org/10.1177/026553229801500101

Runnels, J. (2016). Self-assessment accuracy: Correlations between Japanese English learners' self-assessment on the CEFR-Japan's can do statements and scores on the TOEIC®. *Taiwan Journal of TESOL, 13*(1) 105–137.

Strahan, R. F. (1982). Assessing magnitude of effect from rank-order correlation coefficients. *Educational and Psychological Measurement, 42,* 763-765. https://doi.org/10.1177%2F001316448204200306

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon.

Tokeshi, M., & Gao, L. (2015). CEFR-J based survey for Japanese university students. *The European Conference on Education 2015: Official Conference Proceedings.* http://papers.iafor.org/wp-content/uploads/papers/ece2015/ECE2015_14981.pdf

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

Tono, Y. (2017). The CEFR-J and its impact on English language teaching in Japan. *JACET Selected Papers, 4,* 31–52. https:jacet.org/SelectedPapers/JACET55_2016_ SP_4.pdf#page=41

Walker, D. A. (2003). JMASM9: Converting Kendall's Tau For Correlational Or Meta-Analytic Analyses. *Journal of Modern Applied Statistical Methods, 2*(2) 525-530. https://doi.org/10.22237/jmasm/1067646360

Appendix A

TOEIC Listening and Reading Descriptives (N = 53)

		Listening	Reading (SqRt)
M		314.06	256.13 (15.82)
95% CIs	Lower	296.04	237.45 (15.26)
	Upper	332.10	274.81 (16.37)
5% Trimmed		311.92	252.87 (15.82)
Median		310	255 (15.97)
SE of *M*		9.20	9.53 (0.28)
Variance		4484.67	4815.04 (4.21)
SD		66.968	69.39 (2.05)
Minimum		195	135 (11.62)
Maximum		465	450 (20)
Range		270	315 (8.38)
IQR		90	75 (2.37)
Skewness		0.41	0.56 (-0.06)
Kurtosis		-0.48	0.66 (-0.22)
Shapiro-Wilk *W*		0.97	0.96 (.97)
Shapiro-Wilk *p*-value		.21	.06 (.26)

Appendix B

CEFR-J Listening and Reading Descriptives, Groups and Levels (N = 53)

		Groups		Levels			
Listening		A	B	A1	A2	B1	B2
M		3.33	*2.47	3.43	*3.18	*2.78	*2.16
95% CIs	Lower	3.21	2.33	3.32	3.05	2.63	1.99
	Upper	3.45	2.63	3.54	3.32	2.93	2.33
5% Trimmed		3.33	2.49	3.43	3.19	2.78	2.17
Median		3.20	2.63	3.33	3.00	3.00	2.25
SE *M*		0.06	0.08	0.06	0.07	0.08	0.09
Variance		0.18	0.30	0.18	0.25	0.33	0.41
SD		0.43	0.54	0.42	0.50	0.57	0.64
Minimum		2.60	1.13	2.50	2.00	1.25	1.00
Maximum		4.00	3.63	4.00	4.00	4.00	3.25
Range		1.40	2.50	1.50	2.00	2.75	2.25
IQR		0.70	0.88	0.83	0.50	0.50	0.75
Skewness		0.21	-0.36	-0.02	0.09	-0.06	-0.38
Kurtosis		-1.11	-0.37	-1.09	-0.35	0.54	-0.82
Shapiro-Wilk *W*		0.93	0.97	0.92	0.91	0.96	0.93
P-value of *W*		.004	.221	.001	<.01	.04	.006
Reading							
M		3.30	*2.44	3.44	*3.09	2.84	2.03
95% CIs	Lower	3.18	2.30	3.32	2.94	2.70	1.87
	Upper	3.42	2.58	3.56	3.24	2.99	2.19
5% Trimmed		3.32	2.43	3.47	3.11	2.82	2.03
Median		3.30	2.38	3.50	3.00	3.00	2.00
SE *M*		0.06	0.07	0.06	0.08	0.07	0.08
Variance		0.21	0.26	0.19	0.32	0.29	0.36
SD		0.46	0.51	0.43	0.57	0.54	0.60
Minimum		2.20	1.50	2.50	1.50	1.75	1.00
Maximum		4.00	3.63	4.00	4.00	4.00	3.25
Range		1.80	2.13	1.50	2.50	2.25	2.25
IQR		0.70	0.88	0.83	0.75	0.50	0.75
Skewness		-0.09	0.08	-0.30	-0.06	0.16	-0.12
Kurtosis		-0.60	-0.59	-0.96	0.05	0.27	-0.70
Shapiro-Wilk *W*		0.95	0.98	0.91	0.92	0.94	0.95
P-value of *W*		.048	.494	<.001	.001	0.013	.033

Note. Groups = A (k = 10) and B (k = 8) per skill; and Levels = A1 (k = 6), A2 (k = 4), B1 (k = 4), B2 (k = 4) per skill. "*" indicates hierarchically ordered and significantly different adjacent means.

Appendix C

CEFR-J Listening and Reading Sublevel Descriptives for (N = 53)

Listening		A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2
M		3.68	*3.32	3.29	3.23	3.13	2.96	*2.60	2.39	*1.93
95% CIs	Lower	3.56	3.18	3.15	3.07	2.99	2.80	2.43	2.20	1.76
	Upper	3.80	3.46	3.43	3.39	3.27	3.12	2.77	2.58	2.11
5% Trimmed		3.70	3.34	3.31	3.27	3.15	2.97	2.60	2.39	1.92
Median		4.00	3.00	3.00	3.00	3.00	3.00	2.50	2.50	2.00
SE *M*		0.06	0.07	0.07	0.08	0.07	0.08	0.09	0.10	0.09
Variance		0.18	0.28	0.26	0.34	0.27	0.34	0.40	0.52	0.50
SD		0.43	0.53	0.51	0.59	0.52	0.58	0.63	0.72	0.65
Minimum		3.00	2.00	2.00	1.50	2.00	1.50	1.00	1.00	1.00
Maximum		4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.00
Range		1.00	2.00	2.00	2.50	2.00	2.50	3.00	3.00	2.00
IQR		0.50	1.00	1.00	0.50	0.50	0.50	1.00	1.00	0.50
Skewness		-0.78	0.07	0.15	-0.44	0.08	0.08	-0.11	-0.35	0.08
Kurtosis		-1.19	-0.89	-0.72	0.38	0.05	0.10	0.61	-0.42	-0.79
Shapiro-Wilk *W*		0.68	0.80	0.81	0.87	0.85	0.90	0.89	0.92	0.87
P-value of *W*		<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Reading										
M		3.53	3.41	3.40	3.17	3.02	2.81	2.86	*2.09	1.98
95% CIs	Lower	3.41	3.27	3.27	2.30	2.87	2.65	2.70	1.90	1.82
	Upper	3.65	3.55	3.54	3.35	3.17	2.97	3.02	2.280	2.14
5% Trimmed		3.54	3.44	3.42	3.20	3.03	2.82	2.85	2.06	1.980
Median		3.50	3.50	3.50	3.00	3.00	3.00	3.00	2.00	2.00
SE *M*		0.06	0.07	0.07	0.09	0.08	0.08	0.08	0.10	0.08
Variance		0.21	0.26	0.25	0.42	0.33	0.38	0.35	0.49	0.38
SD		0.45	0.51	0.50	0.65	0.57	0.61	0.59	0.70	0.61
Minimum		2.50	2.00	2.00	1.50	1.50	1.00	1.50	1.00	1.00
Maximum		4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.50	3.00
Range		1.50	2.00	2.00	2.50	2.50	3.00	2.50	2.50	2.00
IQR		1.00	1.00	1.00	1.00	1.00	0.50	0.50	1.00	1.00

Listening	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2
Skewness	-0.28	-0.28	-0.15	-0.40	0.08	-0.39	-0.07	0.21	-0.19
Kurtosis	-1.38	-0.61	-0.69	-0.21	0.09	0.79	-0.28	-0.63	-0.62
Shapiro-Wilk *W*	0.80	0.84	0.81	0.87	0.90	0.90	0.91	0.93	0.88
P-value of *W*	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

Note. Sublevels = A1.1-B2.2 (k = 2) per sublevel per skill. "*" indicates hierarchically ordered and significantly different adjacent sublevels.

Appendix D

Conover Post Hoc and Wilcoxon T-Test Pairwise Comparisons, CEFR-J Levels (N = 53)

	Conover T-Statistic		Wilcoxon T-Statistic			95% CIs	
	(df = 156)	p_{holm}	(df = 52)	p	Effect Size	Lower	Upper
Listening							
A1-A2	3.05	0.003	759.50	<.001	0.85	0.72	0.93
A2-B1	3.77	<.001	697	<.001	0.98	0.97	0.99
B1-B2	3.77	<.001	1046	<.001	0.94	0.88	0.97
Reading							
A1-A2	3.93	<.001	760	<.001	0.95	0.90	0.98
A2-B1	1.64	<.102	561	<.001	0.69	0.43	0.84
B1-B2	5.45	<.001	1215	<.001	0.98	0.97	0.99

Note. For the Wilcoxon test, effect size is given by the matched rank biserial correlation.

Appendix E

Conover Post Hoc and Wilcoxon T-Test Pairwise Comparisons, CEFR-J Sublevels (N = 53)

	Conover T-Statistic		Wilcoxon T-Statistic			95% CIs	
	(df = 416))	p_{holm}	(df = 52))	*p*	Effect Size	Lower	Upper
Listening							
A1.1-A1.2	2.76	0.07	344.50	<.001	0.96	0.93	0.98
A1.2-A1.3	0.38	1.00	108.50	0.578	0.14	-0.35	0.58
A1.3-A2.1	0.42	1.00	106	0.362	0.24	-0.28	0.65
A2.1-A2.2	1.13	1.00	109	0.120	0.43	-0.09	0.76
A2.2-B1.1	1.59	0.79	245	0.023	0.51	0.11	0.77
B1.1-B1.2	2.96	0.04	378	<.001	1.00	1.00	1.00
B1.2-B2.1	1.13	1.00	368	0.016	0.48	0.12	0.73
B2.1-B2.2	2.34	0.18	544	<.001	0.94	0.87	0.97
Reading							
A1.1-A1.2	1.10	1.00	143	0.046	0.51	0.04	0.79
A1.2-A1.3	0.06	1.00	63.50	0.854	0.06	-0.48	0.56
A1.3-A2.1	2.13	0.31	253.50	0.002	0.69	0.37	0.86
A2.1-A2.2	1.44	1.00	243	0.023	0.50	0.10	0.76
A2.2-B1.1	1.66	0.79	287	0.003	0.64	0.31	0.83
B1.1-B1.2	0.81	1.00	145	0.639	-0.11	-0.50	0.33
B1.2-B2.1	5.55	<.001	895	<.001	0.98	0.97	0.99
B2.1-B2.2	0.355	1.00	309.50	0.201	0.25	-0.15	0.58

Note. For the Wilcoxon test, effect size is given by the matched rank biserial correlation. 95% CIs are shown with paired sample Wilcoxon t-tests.

Appendix F

Kendall's τ Coefficients for TOEIC and CEFR-J (N = 53)

				Listening τ	Reading τ
	Groups	Level	Sublevel	(95% CIs)	
All (*k* = 18)				.197 (-.013, .407)*	.322 (.120, .524)***
	A (*k* = 10)			.153 (-.049, .355)	.320 (.131, .509)**
	B (*k* = 8)			.144 (-.065, .352)	.233 (.038, .428)*
		A1 (*k* = 6)		.203 (.017, .389)*	.284 (.107, .462)**
		A2 (*k* = 4)		.108 (-.090, .306)	.294 (.098, .491)**
		B1 (*k* = 4)		.243 (.042, .443)*	.293 (.087, .499)**
		B2 (*k* = 4)		.061 (-.157, .279)	.151 (-.035, .337)
			A1.1 (*k* = 2)	.059 (-.100, .218)	.289 (.125, .452)**
			A1.2 (*k* = 2)	.184 (.011, .357)	.272 (.089, .456)*
			A1.3 (*k* = 2)	.350 (.183, .517)**	.295 (.117, .472)**
			A2.1 (*k* = 2)	.124 (-.067, .314)	.189 (.002, .376)
			A2.2 (*k* = 2)	.038 (-.141, .217)	.374 (.194, .554)***
			B1.1 (*k* = 2)	.191 (.006, .375)	.256 (.074, .438)*
			B1.2 (*k* = 2)	.257 (.051, .452)*	.266* (.082, .450)
			B2.1 (*k* = 2)	.072 (-.140, .283)	.201 (.028, .374)
			B2.2 (*k* = 2)	.057 (-.136, .251)	.064 (-.130, .259)

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.