Preliminary validation of the A1 and A2 sub-levels of the CEFR-J

Judith Runnels
jrunnels@h-bunkyo.ac.jp
Hiroshima Bunkyo Women's University

Abstract

The newly released Common European Framework of Reference Japan (CEFR-J) was designed to address the issue that a consistent system for measuring learner proficiency and progress in foreign language pedagogy in Japan is lacking. This tailored version of the Common Europe Framework of Reference (CEFR) was developed to better discriminate incremental differences in proficiency for Japanese learners of English, who tend to fall mostly within the A1 and A2 levels. Changes from the original CEFR included the creation of can-do illustrative descriptors that separated 4 of the existing 6 levels into sub-levels. The goal of the current analysis is to test the suitability of the new sub-levels of A1 and A2 for target users of the system in two ways: 1) by determining if newly developed descriptors are empirically rank ordered by difficulty as specified by the CEFR-J, and 2) by testing the statistical significance of differences in difficulty ratings between the sub-levels. The current analysis found that the rank ordering of levels was the same as predicted by the CEFR-J, and that the higher-order A1 and A2 levels varied in difficulty to a statistically significant degree, but significant differences between adjacent CEFR-J sub-levels were not found. This raises questions about how users of the system can effectively distinguish features representative of each level and whether the additional sub-levels in the CEFR-J can function as intended. Limitations of using a system of illustrative descriptors based primarily on estimates of difficulty and the process of contextualizing a generalized framework are discussed.

Keywords: Common European Framework of Reference, CEFR-J, can-do statements, difficulty, contextualization

Theoretical work, case-studies and other evidence have suggested that the Common European Framework of Reference (CEFR) provides an effective scheme for describing the needs and outcomes of study for language learners (Council of Europe, 2001). 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. The Framework also defines levels of proficiency which allow learners' progress to be measured at each stage of learning" (Council of Europe, 2001, p. 1). Some argue that the CEFR "is now accepted as the international standard for language teaching and learning" (North, Ortega, & Sheehan, 2010, p. 6).

The framework operates via illustrative descriptors, often referred to as can-do statements, that act to describe what learners are capable of at any given point in time (North, 2007). The CEFR's can-do statements were developed using both qualitative and quantitative methodologies for each level (North, 2000; North & Schneider, 1998). They represent a communicative scheme that gradually progresses from easy to more difficult and are worded in positive terms (Trim, 1978), such that each statement provides a self-sufficient criterion which allows it to be defined independently from other descriptors (Skehan, 1984). The can-do statements are divided into six proficiency levels ranging from Basic User (levels A1 and A2), Independent User (B1 and B2), to Proficient User (C1 and C2) for five skills (listening, reading, spoken production, spoken interaction and writing). Their ultimate aim is to provide a set of learner-centered, performance-related scales which allow for standardized assessment of level (North, 2007).

However, it is in the area of level assessment that the CEFR's suitability and usefulness are frequently questioned and most heavily criticized, particularly with regards to how the can-do statements should be used for test design and evaluation (Weir, 2005). Weir (2005) cautions that in its current form, the CEFR is neither "comprehensive, coherent or transparent for uncritical use in language testing" (p. 281). One issue is related to defining what is entailed by the notion of can-do mastery, a conceptual problem that exists for any system employing illustrative descriptors. As North (2007, p. 13) writes, "what exactly do we mean by 'can do'? Should it be certain that the person will always succeed perfectly on the task? This would be too stringent a requirement. On the other hand, a 50 per cent chance of succeeding would be too low to count as mastery." In order for a system of descriptors to be effectively useable, a definition of mastery is required that is described in terms of how likely it is that a person at a certain level can succeed at a task specified by the can-do - this is one aspect that is sorely lacking from the CEFR. Further criticisms relate to the absence of any theoretical basis in or demonstrable link to work in second-language acquisition (Hulstijn, 2007) when other frameworks, such as the guidelines suggested by the American Council on the Teaching of Foreign Languages (ACTFL) have made this a priority for the last few decades (see Pienemann, Johnston, & Brindley, 1988). Furthermore, there is growing evidence that uninformed usage of the CEFR is leading to assumptions that the CEFR's scales directly tie to stages of language acquisition or specific levels in tests such as the Test of English for International Communication (TOEIC) when in fact, it is derived primarily from difficulty judgments made by language educators (Fulcher, 2003, 2010). Additionally, there is little to no evidence to support the CEFR's pedagogic arguments for gradual development across levels: the hierarchy of difficulty and uni-dimensional or linear progression from easy to more difficult entailed by the framework remains largely unsupported by empirical evidence of performance samples (Westhoff, 2007). Finally, as a basis for test development or any other measures of proficiency, the CEFR and its derivatives do not provide sufficient guidelines for the development of standardized assessments since they lack the information required for either generating test specifications, or being "a medium by which existing tests and specifications can be compared" (Fulcher, 2010, p. 19; see also Fulcher, 2004). Even North (2002), a coauthor of the framework, warns against its usage without full comprehension of its limitations. Ultimately, it is frequently noted by supporters and non-supporters alike, that as long as the CEFR is seen only as a heuristic model and that its limitations are kept in mind, it can nonetheless be employed as a practical and useful tool in constructing curricula, materials and assessments (Fulcher, 2010; North & Schneider, 1998).

In Japan in particular, there is currently no consistently used system for the measurement of achievement of English language learners. Negishi (2011) describes an urgent need for the introduction of a common language framework in Japan in order to start moving towards the widespread, consistent usage of a standardized system for foreign language learning, teaching and assessment. Others have argued the benefits of applying such a system, and specifically the CEFR, to pedagogy in Japan (O'Dwyer & Nagai, 2011). The CEFR was selected as a suitable outlet and serious research on the implementation of the CEFR to foreign language pedagogy in Japan began in 2008 at the Tokyo University of Foreign Studies (Negishi, Takada & Tono, 2011). Difficulty surveys of can-do statements from the DIALANG self-assessment statements (Council of Europe, 2001, pp. 231-234) were administered to 360 Japanese university students. Since they ordered consistently with the CEFR's rank ordering of difficulty, it was concluded that the system was applicable to Japanese learners. Additional findings also demonstrated that over 80% of language learners in Japan skewed towards the A and B levels of the scale (Negishi, 2011). It was concluded that the can-dos across these two levels neither effectively distinguished nor adequately accounted for the variation of ability of language users and development of an alternative version was announced (Negishi, 2011).

Known as the Common European Framework of Reference Japan (CEFR-J), this new version encompasses the following modifications from the CEFR (Tono & Negishi, 2012):

- addition of a Pre-A1 level
- division of A1 into three levels: A1.1, A1.2, A1.3
- division of A2 into two levels: A2.1, A2.2
- division of B1 into two levels: B1.1, B1.2
- division of B2 into two levels: B2.1, B2.2
- adapted can-do statements to a Japanese context

There is a dire need for empirical support of these new statements and level divisions prior to widespread implementation in pedagogy in Japan. As there is currently little research to draw upon from within a Japanese context, the current study was designed to establish a starting point for further research on the newly developed level divisions, or sub-levels, of the CEFR-J at the A1 and A2 levels. Using fabricated or contextualized can-do descriptors has been argued to raise a fundamental question of validity: can a framework function both as a generic reference point and also as a specific application in a local context (North, 2007)? In other words, does the CEFR, which was largely developed and researched within a European context, remain a useable pedagogical tool following modifications and application to a Japanese context? The current study will address this question in two ways: 1) by testing the rank ordering of the can-do statements to determine consistency with the CEFR-J, and 2) by determining if the difference in difficulties between the sub-divisions of levels and categorization of can-dos into each level are statistically significant. Since the CEFR illustrative descriptors have empirically supported interpretations of difficulty (represented in their levels), these difficulty levels should remain consistent if the system is to remain applicable to language regions or educational sectors differing in circumstances to the initial location of development (North, 2007). The first hypothesis is therefore that participants of the current study (target users of the system) will order the can-do statements in the same way as specified by the CEFR-J. Disordered levels would represent a lack of the progression of difficulty entailed by the levels of the CEFR-J and question the underlying assumptions of the system. Secondly, since production of a scale is only the first step in the implementation of a framework, ensuring a common interpretation through empirical support is necessary (North & Schneider, 1998). This requires the existence and identification of features which distinguish one level from the next, or in other words, differences between the estimated difficulties of the newly developed levels. The second hypothesis is therefore that the measures of difficulty across sub-levels will differ significantly from each other. Lack of differences in difficulty would question the thresholds of performance or ability or the features of language required for distinguishing between levels and could result in inconsistent judgements of proficiency.

Methods

Participants

296 first and 294 second-year students of Hiroshima Bunkyo Women's University participated voluntarily in this study. Participants were in one of five disciplines of study: Early Childhood Education, Welfare, Nutrition, Psychology and Global Communication. The survey was administered at the end of July 2012, meaning that the former four major students had completed at least one or three semesters of twice weekly 90 minute university level English classes. The Global Communication majors (a total of 12.5% of participants) had completed either one or three semesters of full-time English study.

Instrument

The survey was administered online using www.surveymonkey.com (SurveyMonkey.com, 2012). Participants were required to indicate the extent of their agreement on a 5-point Likert scale to all 50 Japanese can-do statements for all five skills from levels A1.1 to A2.2. The statements were presented in a random order. These levels were selected because they are the target levels for the institution's curriculum.

Procedure

Since each CEFR-J level is divided into two statements for each of the five skills, there are 10 statements for each level. Due to this being a preliminary investigation, the mean difficulty was calculated for all statements across all skills for each level. The Rasch-measurement software package Winsteps (Linacre, 2010) and PASW Statistics 18 were employed for analysis. The mean Rasch measure, in logits, was calculated for each of the CEFR-J levels from A1.1 to A2.2. Difficulty comparisons across levels were carried out in two ways: first, by measuring differences in the mean logit ratings for each level (where a logit difference of 0.3 represents a significant main effect for difficulty; Lange, Greyson, Houran, 2004; Miller, Rotou, & Twing, 2004) and second, with an ANOVA followed by a least significant difference (LSD) post-hoc test.

Results

Descriptive statistics for the items are shown in Table 1, where a lower logit score represents a lower rating of difficulty. It can be seen that A1.1 had the lowest difficulty rating and A2.2 had the highest, with the remaining levels proceeding in ascending order. The item sub-levels did indeed order by difficulty as hypothesized, although the mean difficulties for each level were very close to each other. This is also evident in Figure 1, which shows a Rasch pathway for the CEFR-J levels. In Figure 1, each level is represented with a circle, whose size is proportional to the standard deviation of the measure for that level. Infit-mean squares are shown on the x-axis.

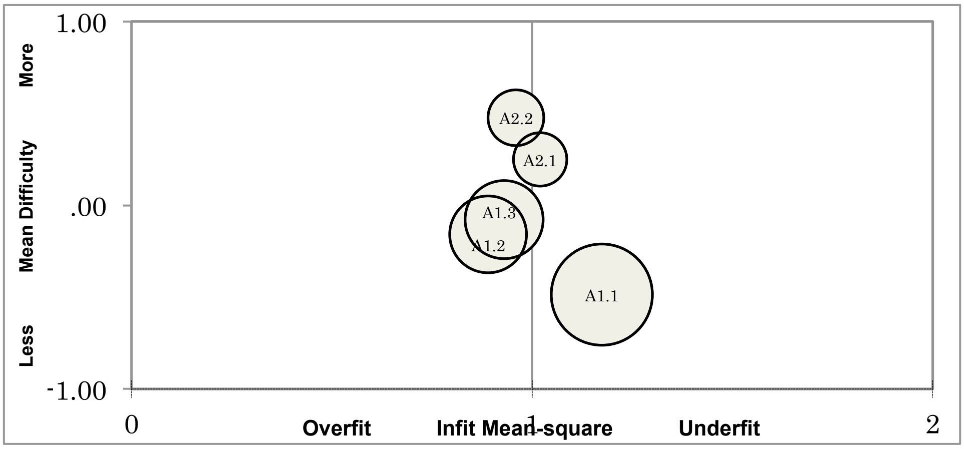

Figure 1. **The bubble chart for CEFR-J levels A1.1 to A2.2.**

Table 1. ***Descriptive Statistics for CEFR-J Levels***

CEFR-J Level	Mean Difficulty	S.D.
A1.1	-0.49	0.584
A1.2	-0.16	0.443
A1.3	-0.08	0.451
A2.1	0.25	0.308
A2.2	0.48	0.322

None of the items exhibited mean-squares outside of the 0.7 – 1.2 range deemed acceptable by Wright and Linacre (1994) and fit statistics for the items have therefore been omitted. As shown in Table 1, the difference in difficulty between levels exceeds the 0.3 logit difference required for significance, for levels A1.1 and A1.2 (0.33) and between A1.3 and A2.1 (0.33). The required difference of 0.3 logits for significance between levels A1.2 and A1.3, or A2.1 and A2.2 was not found. An ANOVA was performed to examine the relationships between levels in more detail. Although differences in difficulties between levels were significant overall ($p = <0.001$; $R^2 = 0.396$), an LSD post-hoc test revealed that there were no significant differences between any adjacent sub-levels (Table 2).

Table 2. ***LSD Post-Hoc Tests for Adjacent Categories***

		Mean Difference	Std. Error	Sig.
A1.1	**A1.2**	-0.33	0.194	0.096
A1.2	**A1.1**	0.33	0.194	0.096
	A1.3	-0.08	0.194	0.693
A1.3	**A1.2**	0.08	0.194	0.693
	A2.1	-0.33	0.194	0.098
A2.1	**A1.3**	0.33	0.194	0.098
	A2.2	-0.23	0.194	0.246
A2.2	**A2.1**	0.23	0.194	0.246

Interestingly, when items were grouped by the original A1 and A2 categories of the CEFR itself, rather than using the sub-levels of the CEFR-J, a statistically significant difference was found. In this case, the overall A1 mean difficulty was -0.24 and the overall A2 mean was 0.36, for a difference of 0.6 logits ($t = 5.075$; $p = <0.001$).

Discussion

The analyses herein were designed to provide empirical evidence on the difficulty of the newly developed levels of the CEFR-J. The first hypothesis tested the rank ordering of difficulty of CEFR-J level statements. The results indicated that the participants ranked the difficulties of the sub-levels in the same way as specified by the CEFR-J (Figure 1 and Table 1). This is not surprising given the extensive process undertaken to create the CEFR-J's can-do descriptors (see Negishi, 2011). Furthermore, previous studies have demonstrated that high correlations between the rank ordering of the difficulty of fabricated descriptors are common (see Jones, 2002; Kaftandjieva & Takala, 2002). Nonetheless, this finding is only preliminary as it compared solely the overall mean difficulty of the sub-levels.

The second hypothesis that differences in difficulty across levels would exist was, unlike the first hypothesis, not supported. Not only did levels A1.2 and A1.3, as well as A2.1 and A2.2 lack the logit difference of 0.3 considered necessary for a main effect of difficulty, but the differences between the remaining levels (A1.1 and A1.2; A1.3 and A2.1), only just meet the threshold of 0.3 logits. When more specific testing was performed using ANOVAs, no significant differences were found between adjacent CEFR-J levels. This raises questions about the division of level A1 and A2 into three and two sub-levels respectively, since the ratings made by users of the system indicate that there seems to be very little to distinguish features representative of these divisions. After reverting back to the divisions of the original CEFR, however, the difference in mean difficulty between the higher order levels of A1 and A2 was both larger and statistically significant. This suggests that the proposed sub-levels for A1 and A2 in the CEFR-J may attempt to make a finer distinction in proficiency than is realistically possible. This will likely represent a challenge for those attempting to place users within the A1 and A2 range, which is also where the majority of Japanese users are purported to lie (Negishi, 2011). As is discussed in Council of Europe (2001, p. 21): "the number of levels adopted should be adequate to show progression…but should not exceed the number of levels between which people are capable of making reasonably consistent distinctions." A potential solution may be to reduce the A1 sub-divisions from three to two and perhaps even A2 into a single level. The same situation may also exist for the sub-divisions of the B1 and B2 levels: further research on this is required to determine if this is a possibility. In either case, this relates back to criticisms of the CEFR: that the assumptions inherent in the hierarchy of levels require supporting empirical evidence (Westhoff, 2007).

A major drawback to the current results however, is that difficulty ratings were averaged across the entire CEFR-J level such that the difficulty was not broken down into separate skills. The data presented herein represent the mean for the entire level across all of the five skills. Future studies should aim to measure the equivalencies between the two can-do statements for each skill for each level, and also across the separate skills. Doing so would better ensure a gradual progression of difficulty across the levels,.

Further limitations of the current study relate to the usage of self-assessment data. It is possible that participants' estimations of whether they have mastered material implicated by the can-do statements are inaccurate: no controls for ability have been employed herein (although this possibility also exists for the participants who were involved in the creation of the system initially—see Negishi, 2011). Investigating this would require comparisons of ability or proficiency derived through other forms of assessment to ensure that more abled students are agreeing with their achievement of the can-dos at higher rates than their lesser abled counterparts. Indeed, one of the criticisms of the CEFR is hinged on this same aspect: that while self-assessment by a language learner or assessment by a language teacher produces scales which order consistently in difficulty between these groups, these scales lack reification (Fulcher, 2010). In other words, it is insufficient for a language framework, if it is to be called that, to be solely based in difficulty estimations by its users (either students and/or teachers), particularly if, as the results in the existing study seem to suggest, the users' behavior does not consistently match predictions by the system.

Conclusion

The results indicated that the participants ranked the difficulties of the sub-levels in the same way as specified by the CEFR-J, though in many cases differences in difficulty between adjacent sub-levels were negligible, and below the threshold of 0.3 logits considered to represent a main effect of difficulty (Lange, Greyson, Houran, 2004; Miller, Rotou, & Twing, 2004).

As Trim (1996) notes, the CEFR deliberately lacks details for local decision-making and action. While it can certainly guide characterizations of language use and language pedagogy, it should not be employed or interpreted as a standardized, benchmarked system. As Davies (2008) points out, when large-scale operations are perceived as ‘the system’, this has historically resulted in a reduction of diversity and experimentation in research surrounding language pedagogy. Ultimately, the investigation of the process of contextualization of the general CEFR to the more local CEFR-J, has revealed that the preliminary work by Negishi, Takada, & Tono, (2011) was relatively successful. The levels ordered as specified and minor differences between some sub-levels were found when target users of the system provided difficulty ratings, although more evidence is required. The results of the present analysis are only a starting point for further validation studies of the CEFR-J, as they do not make any measurements across the language skills or statements contained at each level nor do they provide any controls for proficiency. Establishing the validity of a specialized system which has been developed from a generic reference point is a challenging endeavor (North, 2007). Development alone does not ensure an effective system of measurement or assessment that is capable of specifying the needs, materials or outcomes of study. Research on the CEFR has spanned over twenty years and is ongoing, with continual updates and modifications: the same is required for the CEFR-J. For quality assurance, the system needs to be subject to empirical testing for applicability and effectiveness at every level prior to full implementation or widespread usage in foreign language pedagogy of institutions in Japan.

References

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

Davies A. (2008). Ethics and professionalism. In E. Shohamy (Ed.), *Language testing and assessment.* (pp. 429-443). New York: Springer.

Fulcher G. (2003). *Testing second language speaking.* London: Longman/Pearson.

Fulcher G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266.

Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. *Advances in Research on Language Acquisition and Teaching: Selected Papers,* 15-26.

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

Jones, N. (2002). Relating the ALTE Framework to the Common European Framework of Reference. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* (pp. 167-183). Strasbourg: Council of Europe.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe Scales of Language Proficiency: A validation study. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* (pp. 106-129). Strasbourg: Council of Europe.

Lange, R., Greyson, B., & Houran, J. (2004). A Rasch scaling validation of a ‘core’ near-death experience. *British Journal of Psychology, 95,* 161–177.

Linacre, J. M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Miller, G. E., Rotou, O., & Twing, J. S. (2004). Evaluation of the .3 logits screening criterion in common item equating. *Journal of Applied Measurement, 5*(2), 172-177.

Negishi, M. (2011). CEFR-J Kaihatsu no Keii [The Development Process of the CEFR-J]. *ARCLE Review, 5*(3), 37-52.

Negishi, M., Takada, T., & Tono, Y. (2011). A progress report on the development of the CEFR-J. *Association of Language Testers in Europe Conference.* Retrieved August 1st from: http://www.alte.org/2011/presentations/pdf/negishi.pdf.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

North, B. (2002). Developing descriptor scales of language proficiency for the CEF common reference levels. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment. Case studies.* (pp. 87-105). Strasbourg: Council of Europe.

North, B. (2007). The CEFR Common Reference Levels: Validated reference points and local strategies. *Language Policy Forum Report,* 19-29.

North, B., Ortega, A., & Sheehan, S. (2010). A core inventory for general English, British Council/EAQUALS. Retrieved August 3rd from: http://www.teachingenglish.org.uk/publications/british-council-eaquals-core-inventory-general-english.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–262.

O'Dwyer, F., & Nagai, N. (2011). The actual and potential impacts of the CEFR on language education in Japan. *Synergies Europe, 6,* 141-152.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *SSLA,* 10, 217-243.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*(2), 202–220.

SurveyMonkey.com. (2012). SurveyMonkey. From http://www.surveymonkey.com/

Tono, Y., & Negishi, M. (2012). The CEFR-J: Adapting the CEFR for English language teaching in Japan. *Framework & Language Portfolio SIG Newsletter, 8,* 5-12.

Trim, J. L. M. (1978). Some possible lines of development of an overall structure for a European unit/credit scheme for foreign language learning by adults. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* Strasbourg: Council of Europe, Appendix B.

Trim J. L. M. (1997). The proposed Common European Framework for the description of language learning, teaching and assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma, (Eds), *Current developments and alternatives in language assessment. Proceedings of the LTRC,* (pp. 415-421). Jyvaskyla: University of Jyvaskyla Press.

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

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal, 91*(4), 676 – 679.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.