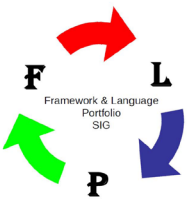
Student ability, self-assessment, and teacher assessment on the CEFR-J's can-do statements

Judith Runnels
Hiroshima Bunkyo Women's University

The Common European Framework of Reference-Japan (CEFR-J), like its original counterpart, the CEFR, uses illustrative descriptors (can-do statements) that describe communicative competencies to measure learner proficiency and progress. Language learners are leveled in a CEFR-J category according to achievement on can-do statements gauged by self-assessment, an external rater (such as a teacher), or from external test scores. The CEFR-J, unlike the CEFR, currently lacks widely-available benchmarked performance samples for measuring student language proficiency, leaving administrations or teachers to estimate CEFR-J ability from test scores or from interactions with students. The current analysis measured ability scores from students and teachers on CEFR-J can-do statement achievement, comparing them to scores on an in-house designed placement test. Students' self-assessment ratings did not correlate with their test scores, teachers varied in severity when making ability estimates for the same students, and no consistent response patterning between students and teachers was found. The results highlight that norming raters, controlling for severity, and training students on self-assessment are likely all required if the CEFR-J is to be used for measuring language learning progress, especially until established guidelines for estimating ability are available for the CEFR-J. The limitations of using the CEFR-J as an assessment tool and the assumption that teachers can accurately estimate student ability are discussed.

ヨーロッパ言語共通参照枠（CEFR）をベースに構築されたCEFR-Japan(CEFR-J)は、学習者の到達度と伸びを測ることを目的に日本の教育機関で最近採用されるようになったシステムである。CEFR-Jは、その基となった枠組みと同様に、段階的に上がる難易度を基にしたコミュニケーション能力を説明するdescriptor（can-doという能力記述文：can-do statements）により構成されている。言語学習者はこのdescriptorの到達度によってレベル分けされる。この評価は、学習者の自己評価、教師などの他の評価者による評価、外部試験の結果から導き出されるものである。これらの評価により、学習者のCEFR-Jにおけるレベルが分かり、標準的にできるであろうとされる能力が示されることになるが、それを使用する人や教師次第になっている部分もある。そこで、もしこのようなシステムを利用する目的が評価レベルの標準化ということであるなら、学習者、教師、そしてテスト評価の判断の間に高い一貫性が保たれなければならない。本論での分析は、CEFR-Jのdescriptorについての学生と教師の能力判断の一貫性、そしてその判断が学内作成のプレイスメントテストの点数と一致するかを検証することを目的としている。学生と教師の判断には顕著な関係はみられず、学生の自己評価の結果はテストの点数と相関性がなかった。この結果により、もしCEFR-J　が評価の標準化を目的に使用されるのであれば、規範的な評価者と自己評価についての学生指導の必要性が重要になるといえる。評価のツールとしてCEFR-Jを使うことの限界、及び説明的なdescriptorのシステムに本来備わるcan-do熟達度という概念に関する問題を議論する。

SIG Spotlight: FLP SIG



The Framework & Language Portfolio (FLP) SIG wants to discuss the Common European Framework of Reference for Languages (CEFR) and European Language Portfolio (ELP), the related pedagogical implications, and their relevance for language education in Japan while carrying out projects and communicating the results. There is an emphasis on developing materials to support educators who would like to use these pedagogic tools. Our members hold fora at JALT conferences, participate in other events, and engage in research projects. See the FLP SIG Kaken Project <tinyurl.com/FLPKaken> for examples.

Alongside edited volumes, e.g., can-do statements in language education in Japan and beyond, the FLP SIG also publishes a newsletter two to three times a year. Members receive these once published, and back issues are available at the link above. Topics include updates on ongoing projects and events, and generally include a feature article. A summary of one such article is found in this edition.

The Common European Framework of Reference (CEFR) describes the needs, goals, and outcomes of study for language programs and autonomous learners (Council of Europe [COE], 2001). Illustrative descriptors (can-do statements), in six levels of proficiency, describe communicative competencies in listening, reading, spoken production, spoken interaction, and writing (COE, 2001; North, 2000, 2007 & Schneider, 1998). It is argued that the CEFR "allow[s] progress to be measured at each stage of learning" (COE, 2001, p. 1) and provides sets of scales for standardized ability assessments (Little, 2005; North, 2007). Others note that can-

do statements alone do not provide sufficient criteria for proficiency evaluations (Fulcher, 2003, 2010; Weir, 2005).

Since measurements derived from can-do statements are used for measuring proficiency, some consistency between and across the judgments made by the different populations of users (i.e., students, teachers, or other raters) can be expected. Previous research, however, has suggested that teachers are incapable of making accurate judgments on their students' abilities (Bérešová, 2011; North & Jones, 2009), despite the fact that administrations continually require them to do so (Protheroe, 2009). Additionally, very few studies take a learner's self-assessment—one of the most important components for autonomous learning (Holec, 1979; Little, 2006)—into consideration.

The current study was therefore designed to examine judgments of achievement from teachers and students on can-do statements and their relationship with test scores. The can-do statements from the CEFR-Japan (CEFR-J), an alternate version of the CEFR tailored to meet the needs of Japanese learners of English in Japan, were used to measure this relationship (see Negishi, 2011; Negishi, Takada, & Tono, 2011; Tono & Negishi, 2012). Since the CEFR-J was developed at least partly for the purposes of standardized assessment, in order for it to be used as such, the perception or understanding between users of what is required to achieve each level should be somewhat consistent. It is therefore hypothesized that students' self-assessments, test scores, and teachers' assessments should mirror each other to some extent.

Methods

Participants

Participants were 296 first year university students in one of the ten classes streamed for ability by a placement test. Four classes (69 participants) were omitted, being either English majors or the highest scoring individuals on the placement test. Participants were unfamiliar with the CEFR-J and had no prior experience using can-do statements or conducting self-assessments.

Teacher participants consisted of seven native English-speaking staff members who had worked with the ten classes of students throughout one semester of study. All teachers were relatively familiar with the CEFR-J and its can-do statements.

Instruments

Participants indicated the extent of their agreement on a 5-point Likert scale (from *Strongly Disagree* to *Strongly Agree*) to all 50 randomly ordered Japanese can-do statements from the CEFR-J's A sub-levels (A1.1, A1.2, A1.3, A2.1 and A2.2; TUFS Tonolab, 2012).

Teacher participants responded to the same 50 randomly ordered can-do statements in English, indicating to what extent they believed that 80% of their students could perform the can-do statement. Eighty percent was chosen as this threshold is frequently used in domain or criterion-referenced testing as an indication of mastery (North, 2007), and is used as a guideline for teachers to estimate student ability and select appropriately targeted classroom materials (Protheroe, 2009).

The assessment used to control for ability and measure the relationship between ability and self-assessment scores was an in-house designed reading and listening test developed for the purposes of streaming students into leveled classes (Runnels, 2013). It had been administered three months prior to the can-do survey and it should therefore be noted that any gains or losses in proficiency between the times the test and the survey were administered have not been taken into account.

Procedure

Mean achievement ratings on listening and reading can-do statements for all students in each class were compared to the teachers' rating for the class on each skill. It should be noted here that the scores are not expected to match exactly, but if the CEFR-J is to function as intended, similar response patterns between groups are predicted. However, there are significant issues with comparing teacher ratings on an entire group to mean ratings from a group of individuals, although this is precisely what frequently happens in institutions (Protheroe, 2009). Ideally, teachers would rate individuals, but not only was this deemed unreasonably time-consuming, judging students individually has not been found to improve the accuracy of teachers' estimations (Bérešová, 2011).

Student can-do statement self-assessment scores were also correlated with their individual test scores to examine the relationship between self-assessment and ability. Although classes exhibited the same mean score overall, individuals making up the classes naturally varied in their

Table 1. Descriptive statistics for student and teacher ratings on can-do statements

	Mean (*S.D.*)		Range (Minimum – Maximum)	
	Teachers	Students	Teachers (Likert Scale Point Spread)	Students (Likert Scale Point Spread)
Listening	2.59 (*0.44*)	3.5 (*0.15*)	2.2 – 3.2 (1.0)	3.24 – 3.8 (0.56)
Reading	2.80 (*0.55*)	3.4 (*0.17*)	1.8 – 3.7 (1.9)	3.1 – 3.7 (0.6)
Overall	2.63 (*0.52*)	3.45 (*0.16*)		

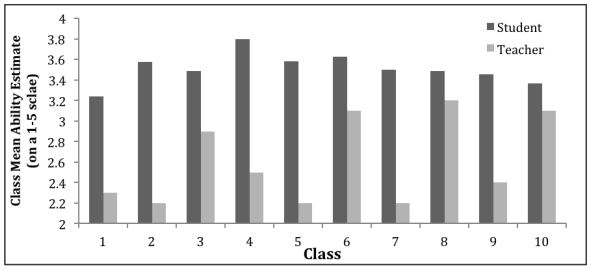


Figure 1. Mean ability estimates for each class on the CEFR-J's A1.1 – A2.2 listening can-dos.

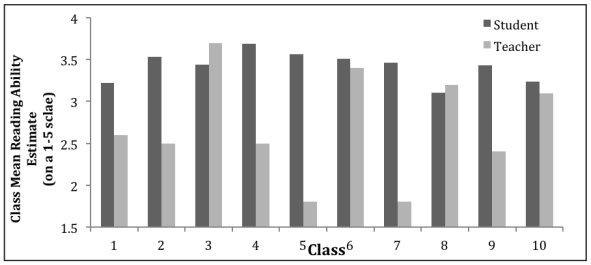


Figure 2. Mean ability estimates for each class on the CEFR-J's A1.1 – A2.2 reading can-dos.

test scores. Since administrations assume overall class abilities to be equal, within-class variance was not accounted for.

Results

Descriptive statistics for both the student and teacher surveys are shown in Table 1. Figures 1 and 2 show the results of the student and teacher surveys for listening and reading for each class.

Despite teachers giving a significantly lower mean achievement rating for students, both the standard deviation and the range of teacher responses are much larger than for the students' self-assessments (Table 1). Additionally, the correlation between students' test-scores and CEFR-J can-do self assessment scores were essentially nil (r = .005): Students' achievement ratings were similar across all classes but did not correlate with their test scores, whereas teachers' ratings differed both from students' judgments and from the ratings of other teachers.

Discussion

The results indicated no consistent relationship between teacher and student judgments on can-do statement achievement. Furthermore, the students' self-assessment scores did not correlate with test scores used to measure ability. Finally, there was little agreement between teachers on student ability. These results raise questions about how can-do statements can be used for standardized assessment if there are such large discrepancies in understanding between teachers and between teachers and students. It also reiterates findings of previous research: There is little evidence to support the assumption that teachers can accurately estimate their students' ability.

These findings highlight several issues regarding self-assessment by Japanese learners, student ability assessment across teachers, and also between teachers and their classes. Regarding the former, Japanese survey-takers in general have been shown to both gravitate toward selecting neutral responses (Dörnyei & Taguchi, 2010) and, for self-assessment surveys in particular, be subject to Japanese cultural factors related to modesty (Matsuno, 2009; Takada & Lampkin, 1996). Japanese students, therefore, likely require significant training in using CEFR-J can-do statements for meaningful self-assessments. In fact, Japanese institutions should perhaps aim to emphasize this in their language programs (there are many resources available for this: Blanche & Merino, 1989; Glover, 2011; Gonzales, 2009; Holec, 1979; Little, 2006; Rolheiser & Ross, 2013; Zhou, 2009).

In terms of the inconsistent judgments on student ability from teachers, this can be attributed to rater-reliability and a lack of controls for rater severity. Without adjustments for rater severity, raw judgment ratings cannot be

directly compared to each other (Wright, 1998) and institutions would be remiss in doing so. Rater training (or norming), which might consist of familiarization to the CEFR-J and the use of can-do statements, followed by workshops on how to create, localize, align, and use can-do statements would ensure higher reliability (Elder, Barkhuizen, Knoch, & von Randow, 2007; Weigle, 1998; Woehr & Huffcutt, 1994) (also see Harsch & Martin, 2012 for CEFR-based rater training). In fact, the COE (2003) offers DVDs of sample performances, illustrating requirements at each CEFR level for English and French, although these resources do not yet exist for the CEFR-J (North, 2007).

The findings presented here also have implications for the usage of the CEFR-J at an institutional level, particularly regarding curriculum planning and materials selection. The current study illustrates disagreement between teachers about students' language ability. The selection of materials or tasks deemed appropriately targeted to students' abilities would thus differ depending on the teacher, and students may not agree that the selected materials are suitable for them. To address this, a tool such as DIALANG (Alderson & Huhta, 2005), which provides proficiency estimates of level based on performance derived from the CEFR's can-do statement-tasks, might be beneficial to both teacher and student users of the CEFR-J in estimating level.

The present findings, though preliminary due to limitations, emphasize nonetheless that a more thorough investigation of the relationship between learner self-assessment, language ability, and assessment by external raters is required for the CEFR-J. If replication studies (ideally with can-do surveys and placement tests being administered at the same time) also show that, despite training, students make more lenient ability judgments than teachers, teachers continue to exhibit substantial ranges of severity in their judgments after adjustments, and that either of these tendencies is inconsistent both within or across groups, the consequences for the CEFR-J are significant. Findings such as these would question how, in its existing form, the CEFR-J can be used as a tool for the assessment of (or for) learning, and administrations should be cautious about making major decisions without further research.

References

Bérešová, J. (2011). The impact of the Common European Framework of Reference on teaching and testing in Central and Eastern European context. *Synergies Europe, 6*, 177-190.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, *39*(3), 313-338.

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

COE. (2003). *Relating language examinations to the Common European Framework of References for languages: Learning, teaching, assessment: Preliminary pilot manual*. Strasbourg, France: COE, Language Policy Division.

Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing*. New York: Routledge.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

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.

Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness*, *20*(2), 121-133.

Gonzalez, J. A. (2009). Promoting student autonomy through the use of the European Language Portfolio. *ELT Journal*, *63*(4), 373–382.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228-250.

Holec, H. (1979). *Autonomy and foreign language learning*. Strasbourg, France: Council of Europe.

Little, D. (2005). The Common European Framework and the European Language Portfolio: involving learners and their judgments in the assessment process. *Language Testing, 22*(3), 321-336.

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*, 167–190.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing. *Language Testing*, *26*(1), 75-100.

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*. Paper presented at the 4th Association of Language Testers in Europe International Conference Retrieved from <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. (2007). The CEFR Common Reference Levels: Validated reference points and local strategies. *Language Policy Forum Report*, 19-29.

North, B., & Jones, N. (2009). *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment (CEFR) further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT Scaling*. Strasbourg, France: Council of Europe.

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

Protheroe, N. (2009). Improving teaching and learning with data-based decisions: Asking the right questions and acting on the answers. Retrieved from <www.lesn.appstate.edu/olson/RES5080/Components/Articles_used_in_5080/Pruthero%20Improving_teaching_and_learning_with_databased_decisions.pdf>

Rolheiser, C., & Ross, J. (2013). Student self-evaluation: What research says and what practice shows. Retrieved from <cdl.org/resource-library/articles/self_eval.php>

Runnels, J. (2013). Evaluation of a streaming instrument. *Kanda University of International Studies Journal*, *25*, 119-131.

TUFS Tonolab. (2012). CEFR based framework for ELT in Japan. Retrieved from <www.tufs.ac.jp/ts/personal/tonolab/cefr-j>

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-267.

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

Woehr, D., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189-205.

Wright, B. D. (1998, September). How to convince your friend not to use raw scores. Paper presented at the COMET Meeting, Institute for Objective Measurement & MESA Psychometric Laboratory.

This article first appeared in the *Framework and Language Portfolio Newsletter*, *9*, pp. 6-18. It has been shortened and edited for the current publication.

Judith Runnels was most recently a lecturer at Hiroshima Bunkyo Women's University. Her research interests include the CEFR-J and the assessment and evaluation of language placement and speaking tests. She has previous teaching experience in Canada, China, and Korea. She can be contacted at <judith.runnels@gmail.com>.