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.

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, 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; TUFSTonolab, 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
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 select 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

<table>
<thead>
<tr>
<th></th>
<th>Mean (S.D.)</th>
<th>Range (Minimum – Maximum)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Teachers</td>
<td>Students</td>
</tr>
<tr>
<td>Listening</td>
<td>2.59 (0.44)</td>
<td>3.5 (0.15)</td>
</tr>
<tr>
<td>Reading</td>
<td>2.80 (0.55)</td>
<td>3.4 (0.17)</td>
</tr>
<tr>
<td>Overall</td>
<td>2.63 (0.52)</td>
<td>3.45 (0.16)</td>
</tr>
</tbody>
</table>
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


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>.

JALT Other Language Educators SIG invites you to the 2nd JALT OLE SIG Conference

LanguageS PLUS

Language learning and teaching beyond the first foreign language
• Date: Oct. 12/13, 2013
• Venue: Chukyo University, Nagoya
• Info: <www.geocities.jp/dlinklist/ENG/OLEkon2013.html>

Early proposal submissions with the subject title OLE2SIG to the OLE Coordinator:
<reinelt.rudolf.my@ehime-u.ac.jp>
T/F 089-927-9359

Please inform teachers of languages other than English and Japanese of this opportunity

Judith Runnels