Taiwan Journal of TESOL
Vol. 13.1, 105-137, 2016

SELF-ASSESSMENT ACCURACY: CORRELATIONS BETWEEN JAPANESE ENGLISH LEARNERS' SELF-ASSESSMENT ON THE CEFR-JAPAN'S CAN DO STATEMENTS AND SCORES ON THE TOEIC®

Judith Runnels[1]

ABSTRACT

Since its release in 1979 the TOEIC® (Test of English for International Communication) has been consistently and widely used by educational institutions and companies of Japan despite criticisms that it provides little useable information about language ability. In order to both reduce the extreme focus on and also aid with the practical interpretability of TOEIC® test scores, other approaches to the assessment of language proficiency have started to gain popularity. One notable shift seems to be towards the usage of the Common European Framework of Reference (CEFR), which is purported to provide a highly learner-centered approach to the teaching, learning and assessment of languages. The CEFR promotes the development of learner autonomy and supports learner self-assessment through the usage of can do statements, which describe the communicative actions learners are able to perform at any given time. Due to the increasing interest in using the CEFR as an assessment tool for learning in Japan, further study of the relationship between language proficiency and self-assessment is required. The current study thus explored the relationship between Japanese English language learners' self-assessment scores on listening and reading can do statements from the Common European Framework of Reference-Japan (CEFR-J, a modified version of the CEFR) with test scores from the TOEIC. Moderate correlations between the TOEIC and can do self-assessment scores were found for listening, while no correlations were found for reading. The factors that may influence a learner's self-assessment tendencies, the efficacy of a self-assessment system for Japanese learners and the interpretability of TOIEC® scores are discussed.

Key Words: TOEIC®, CEFR, CEFR-Japan, self-assessment, language proficiency, can do statements

[1] PhD student, University of Bedfordshire, Luton, UK

BACKGROUND

Since its release in 1979, the listening and reading Test of English for International Communication (TOEIC®, hereafter TOEIC) has continually gained in popularity in Japan as a standardized assessment of the language skills required by Japanese learners of English in international workplaces (Gilfert, 1996; Woodford, 1982). Given that Japanese nationals make up the greatest number of test-takers globally, and that the TOEIC has been used in Japan for over thirty years, it provides a measure of language proficiency that many Japanese institutions are familiar with and frequently enquire about (Chapman, 2003; Childs, 1995; Ito, Kawaguchi, & Ohta, 2005). TOEIC scores are used for the following purposes: evaluating the effectiveness of internally designed language training programs, assessing the English abilities of prospective employees, setting requirements in making decisions about promotions or overseas assignments, or more vaguely, maintaining competitiveness in national or global economic markets (Ito et al., 2005). Despite its widespread usage, it has been noted that the TOEIC may not be an appropriate instrument for any of the aforementioned purposes because the interpretability of its scores is problematic and it is limited in its ability to provide any useable information on the language proficiency of the test-takers (Chapman, 2003; Childs, 1995; Wilson, 1989). Ito et al. (2005, pp. 1-2) pose the following questions: “What does the score mean? How should score recipients interpret their scores? . . . Are companies able to predict what an employee with a score of 600 is capable of doing with English in their working environment?” According to Kubota (2011), the answer to the latter question is ‘No’ and organizations should be cautious to overemphasize the importance of obtaining a certain TOEIC score in order to gain employment, as there may be neither much benefit nor use for such a requirement.

A similar situation prevails in Japanese universities, where learners are sometimes required to obtain a threshold TOEIC score in order to graduate (Shibata & Inoue, 2005). Concerns regarding the impact this has on the educational environment have consequently been raised: Institutions’ internal educational goals are compromised for the sake of the external organisation’s examination goals, educators become forced to engage in a TOEIC score competition with other institutions or even within their own institutions, students focus on their score rather than their English language proficiency, and language skills not tested on the

TOEIC (speaking and writing) are underemphasized (Shibata & Inoue, 2005). This is not to say that language learners do not or cannot benefit from feedback derived from what Gardner (2000) refers to as large-scale institutionalized assessments, just that sole emphasis on that type of assessment does not necessarily represent a holistic or personalized approach to language learning (Gardner & Miller, 1999).

Consequently, some educational institutions are in the midst of shifting away from the focus on TOEIC scores towards what is arguably a more individualized approach to language teaching, learning and assessment (O'Dwyer & Runnels, 2014) where test scores are downgraded in their importance (Nagai & O'Dwyer, 2011). To provide an example specific to Japan, Osaka University uses the Common European Framework of Reference (CEFR; Council of Europe, 2001) to frame the programs and courses of the 25 language degrees it offers (O'Dwyer & Runnels, 2014). Other tertiary level institutions around the country, both private and public, have also referred to the CEFR in the design of their language learning programs (see O'Dwyer, Nagai, Imig, Naganuma, Schmidt, & Hunke, n.d.; Schmidt, Naganuma, O'Dwyer, Imig, & Sakai, 2010). Doing so is argued to create a synergy between the three areas of learning, teaching and assessment, resulting in a positive learning and assessment culture, both within and outside of the classroom (O'Dwyer, Imig, & Nagai, 2014; O'Dwyer & Runnels, 2014). Additionally, many companies are choosing to use other tests of proficiency (Eiken, n.d.) such as BULATS (University of Cambridge Local Examinations Syndicate, 2016), the score of which is reported in CEFR levels, and supported by language proficiency descriptors which describe what the language users are able to do in English.

The CEFR and Self-assessment

The CEFR is the description of the Council of Europe's (CoE) language policy, produced for the purposes of increasing collaboration and cooperation between European educational institutions (Trim, 2007). As a system used to describe communicative language competences, the CEFR intends to be extensive, coherent, and transparent. The CEFR is best known for its descriptors of language proficiency, or can do statements, which are divided into five language sub-skills across six levels (see CoE, 2001, 2005; Little, 2007; Trim, 2007). Since the CEFR's publication, it has impacted foreign language education industries both within and outside of the CoE's member states with many identifying it as "[an] international standard for language teaching and learning" (North, Ortega, & Sheehan, 2010, p. 6). The CEFR is purported to support a number of facets of language education, including the planning of the content, objectives or assessment criteria of language learning programs and language certification, the selection of materials for self-directed learners and for the evaluation of learning or learner progress (CoE, 2001, p. 7). It also intends to provide a set of learner-centered scales which allow for a standardized assessment of proficiency (North, 2007). The CEFR is also criticized though, particularly for its usage in assessment (see Alderson, 2007; Fulcher, 2003, 2004, 2010; Hulstijn, 2007; Weir, 2005). Further critiques relate to the lack of support to the purported progression of difficulty inherent to the framework which is neither tied to stages of language acquisition nor evidenced by empirically obtained performance samples (Westhoff, 2007).

Nonetheless, one of the CEFR's strengths is that its scales of can do statements permit learners to both define their own abilities in their language of study, and plan the direction of their future studies (CoE, 2001; Glover, 2011; Little, 2006), both of which contribute to the development of autonomous learners (O'Dwyer et al., 2014; O'Dwyer & Runnels, 2014). This is typically done through a learner self-assessment using can do statements. A learner may read a statement and then make a decision regarding their perceived performance of the communicative task implied by the statement (Glover, 2011; Little, 2006). To provide some examples, Table 1 shows the CEFR's listening statements for learner self-assessment from levels A1 to C2 (CoE, 2001, pp. 26-27). If the learner believes they can perform sufficiently or proficiently within the area of that can do statement, they may move on to responding to a statement from another skill or a more difficult statement within the same skill. If they reach a statement and feel they are unable to complete the implied task, they would likely perform some further studies in that area until they are more comfortable with their proficiency. In this way, the framework provides the general scaffold for the learner's progress as they gradually focus on higher level can do statements over time.

Table 1

Self-assessment Can Do Statements for Listening from the CEFR's Levels A1 to C2

CEFR Level	Can do statement
A1	I can recognize familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.
A2	I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.
B1	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.
B2	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.
C1	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signaled explicitly. I can understand television programmes and films without too much effort.
C2	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.

Although the CEFR is best known for its can do statements and reference levels shown in Table 1 (CoE, 2005; Martyniuk & Noijons, 2007) and the CEFR's scales of can do statements have also been most often used for learner self-assessment (North, 2007), many questions as to how the usage of can do statements can help learners work towards their learning goals, develop pathways for future study, help with material selection, or achieve any of the other of the CEFR's goals exist since it is not clear if responding to can do statements can truly provide an estimate of language proficiency, if even a general one (Green, 2012). Kodate and Foale (2012, p. 33) add that “simply getting learners to answer ‘Yes, I can.’ or ‘No, I can’t.’ to questions about their language ability does not guarantee that they will utilize can-do statements [or the results of a self-assessment] in a meaningful way”. Even though the CEFR intends to permit learners to measure or estimate language proficiency, neither is the relationship between self-assessment and actual language ability nor is the performance of the CEFR’s can do statements as a self-assessment instrument well-enough understood to be able to provide such a measure (Tavakoli & Ghoorchaei, 2009). The need for the current investigation stems from these concerns, and aims to investigate issues surrounding the general relationship between self-assessment and language proficiency.

Self-assessment

The ability to self-assess is often seen as a keystone characteristic of an autonomous language learner (Gardner, 2000; Holec, 1981; Thomson, 1996), where autonomy refers to a learner’s capacity to take charge of and responsibility for their own learning (Holec, 1981). Self-assessment refers not to the construction or constructor of the assessment, but to the mode of administration in that it is self-administered (Gardner, 2000). Such an assessment may serve a number of purposes, such as measuring progress, proficiency, motivation or confidence (Gardner & Miller, 1999). Self-assessment has been associated with a wide-ranging array of benefits for language learners, but it is also a “technique that needs to be introduced [to learners] carefully and accompanied by considerable awareness raising and support” (Gardner, 2000, p. 49). Moreover, there are a number of concerns related to self-assessment which are significant enough to dissuade any teacher or learner from performing them, perhaps the most obvious of which is the reliability of the results (discussed in the next section). A further drawback is the perceived face-validity of the assessment, which is compromised if stakeholders believe that it is either easy to cheat on, or of no worth in the first place (Gardner, 2000). Gardner (2000, p. 54) also states that self-assessment can “upset the perceived balance of power” in an educational setting: typical attitudes towards assessment suggest that it is seen as a task which should be mediated by the teacher. When this does not occur, learners may interpret the self-assessment to be the result of having a lazy teacher, and resent the teacher for assigning the extra work. Furthermore, learners “may feel unequipped or unwilling to produce, conduct and interpret their own assessments” or be too nervous to try, in the same way that they feel self-conscious about using a foreign language in the first place (Gardner, 2000, p. 54). Despite the pitfalls of self-assessment, which Gardner (2000) argues can be significantly minimized if not entirely mitigated by the teacher, the issue of reliability of self-assessment remains: Can self-assessments provide an accurate representation of language ability?

Self-assessment and Language Proficiency

In terms of the conduct of an investigation into the reliability of self-assessment, Sundstroem (2005) reviewed self-assessment with ability comparisons across a number of fields, concluding that the correlation of self-assessment results with proficiency tests was the most prominent method employed. For language learning this also seems to be the case: the correlation of proficiency with “self-assessment has been [most often] investigated… by means of correlating self-estimated ability data with more objective measures of the same abilities” (Ito et al., 2005, p. 3). Edele, Suering, Kristen, and Stanat (2015) reviewed over 30 correlational studies between language self-assessment and language test scores. The results of such studies have been mixed: Bachman and Palmer (1989), Blanche (1990), Blanche and Merino (1989), Finnie and Meng (2005), and LeBlanc and Painchaud (1985) have all found self-assessment scores to be highly reliable, with the accuracy of students’ self-estimates to be either good or very good. Brantmeier, Vanderplank, and Strube (2012) and Alderson (2005) found only moderate-strength correlations between self-assessment and test scores. Conversely, Thomson (1996), Pierce, Swain, and Hart (1993) and Janssen-van Dieten (1989) noted considerable divergence between learners’ ratings on their ability and their proficiency test scores. Blue (1988) and Runnels (2013a) found very few similarities in proficiency ratings between students’ own estimations of their ability and those of their teachers’. Chen (2008) also compared student and teacher assessment scores and noted that only after significant feedback and practice did the students’ assessments echo those of their teachers. Davidson and Henning (1985, p. 175) warned that the “phenomenon of exaggeration in ability estimation [is] an inherent weakness of self-reports of language abilities".

Building on these findings, studies exploring the factors influencing the accuracy of self-assessment emerged (Ito et al., 2005). Spolsky (1992) for instance, suggests that self-assessment accuracy depends on whether the responses reflect aspects of language proficiency which lie within the experience of the responder. Ross (1998) also found that the extent to which students' self-assessment matched teachers' ratings and test scores was dependent on how much experience the learner had with the language skill being self-assessed. Others have deemed general competence of test-takers, specificity and difficulty of the assessed domain (Sundstroem, 2005), variables external to question content, (Heilenman, 1990) and bias and self-esteem (MacIntyre, Noels, & Clément, 1997) to be influential on the accuracy of self-assessment.

In all of the aforementioned studies however, both the self-assessment battery and the objective assessment instrument varied according to the context. Indeed, the majority of self-assessment and language proficiency studies discussed herein report vastly differing correlations (from zero to very high) which could be due to the variation in quality of the instruments being used for both self-assessment and proficiency measurement (Edele et al., 2015). It was therefore suggested that similar investigations performed using standardized and more widely established and accessible instruments would provide a better understanding of the relationship between self-assessment and language proficiency (Edele et al., 2015).

The TOEIC and the CEFR's can do scales, both widely recognized and established instruments, were therefore selected for use in the current investigation into learner self-assessment and the accuracy of its results. The selection of both the TOEIC and the CEFR's can do statements is also due to their relevance to the language education context in Japan. It is hoped that their usage in this study will provide practical, useable information for anyone interested in using self-assessment within their own student or employee populations (Edele et al., 2015). Gardner (2000) supports such an approach, and recommends that teachers conduct their own research into self-assessment with their learners, and subsequently share the results with both colleagues and learners so that each stakeholder can make their own judgements about the results. The ideas presented by Gardner (2000) and Edele et al. (2015) suggest that self-assessment accuracy could differ across various contexts including between learners in the same context, similar learners in different contexts, or even within an individual learner over time. It is therefore deemed important to consider established characteristics of Japanese self-assessors, as doing so will likely provide a better understanding of the results of their self-assessments.

Japanese Self-assessors

Exploring self-assessment in Japanese learners is a particularly relevant issue in the current educational landscape of Japan given the Japanese Ministry of Education, Culture, Sports, Science and Technology's (MEXT) publication of measures to develop secondary students' proficiency in English by mandating the use of lists of self-assessment can do statements in secondary schools (MEXT, 2013; Takada, 2014). Furthermore, as evidenced by increases in publications and conferences on the framework, the CEFR is also gaining in popularity in Japan (Shimo & Nitta, 2011). Despite the fact that extensive validation studies to confirm the CEFR's hierarchy of difficulty have been performed (North, 2000, 2002; North & Schneider, 1998), it has been suggested that the CEFR's can do statements may not perform as intended when administered to Japanese language learners for the purposes of self-assessment (Runnels, 2013b, 2013c).

Despite extensive findings in the field of self-assessment in general, and even for the self-assessment of language proficiency, for Japanese self-assessors, a comparable lack of work is notable. With the exception of Ross (1998) and Runnels (2013a), none of the studies discussed in the previous section focused on Japanese learners of English. Indeed, research on Japanese survey takers in general has revealed some unique phenomena in the response patterns of participants. For example, Japanese survey-takers tend to select neutral responses no matter the content of the item (Dornyei & Taguchi, 2010), while Japanese self-assessors have been generally shown to respond according to social desirability factors related to the perception and exhibition of modesty (Ikeno, 2002; Matsuno, 2000; Takada & Lampkin, 1996). However, it is unclear if these same phenomena apply to Japanese English learners while self-assessing their language abilities.

Focus of the Study

Given the issues surrounding the usage of the TOEIC and the recent increase in the CEFR’s popularity in Japan, the MEXT’s language learning policies, and the need for further study into the relationship between self-assessment and language proficiency for Japanese self-assessors in particular, the present investigation was conceived. This study aims to determine if higher proficiency in English is also associated with higher self-assessment scores, and also hopes to provide a departure point for future investigations into any of the aforementioned issues, including the usage of TOEIC scores as a measure of language proficiency for Japanese learners of English and the functionality of the CEFR-Japan (CEFR-J) as a self-assessment instrument. The CEFR-J is a localized version of the levels and scales of the CEFR and was specifically created in order to meet the unique demands of the Japanese educational context. The CEFR-J contains modified can do statements (adapted for Japanese learners of English) and a greater number of levels within the CEFR’s global A and B levels (for the development process, see Negishi, 2011; Negishi, Takada, & Tono, 2013; Tono & Negishi, 2012).

The current study therefore explores Japanese university English majors’ self-assessments on the CEFR-J’s listening and reading can do statements and how they compare with scores obtained on the TOEIC. Specifically, TOEIC listening and reading scores were correlated with self-assessment ratings on listening and reading can do statements from the CEFR-J’s nine A and B sub-levels (A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2). The following research questions were investigated:

1. What are the results of English majors’ self-assessments on CEFR-J can do statements?
2. Are the learners’ self-assessments in accordance with the predicted difficulty hierarchy of the CEFR-J (are higher-level can do statements rated as more difficult than the lower-level statements?)?
3. What is the relationship between Japanese English language learners’ TOEIC and CEFR-J self-assessment scores for listening and reading?

METHOD

Participants

A cohort of 80 English majors from a small private university in Japan were the participants in this study. To graduate, there is a soft requirement that students obtain a TOEIC score of 600, although other measures can be taken if students are unable to achieve such a score. This requirement is flexible because the degree program is not geared towards ensuring the learners succeed on the TOEIC. Instead, it follows a CEFR-informed curriculum which uses can do statements from various levels as overall course goals. In order to be eligible for inclusion in the analysis, students had to have taken the TOIEC and completed the can do survey within a period of less than two weeks. For 23 of the 80 English majors, their most recent TOEIC scores were from at least four months prior, and in some cases eight months or more. It was thought that this was too long a time for comparability with the rest of the students, as it was expected that significant improvements in language ability would have occurred throughout those four or more months of full-time English study. Data from 23 participants was therefore removed, leaving 57 first-, second-, and third-year majors as participants.

All participants were familiar with the concept and process of self-assessment through a weekly class which focuses on the introduction and evaluation of learning strategies, of which a major focus is self-assessment. According to the teacher, in this class "learners are introduced to the conceptual ideas and the benefits of independent learning, some of the core skills required to conduct independent learning (e.g. needs analysis, goal setting, materials selection, and evaluation), and strategies for English language learning" (Kodate, 2012, p. 129). Subsequent to the input of the theory segment of the course, "learners carry out their own independent English learning projects, thereby transforming the knowledge they acquired from the course into practice" (Kodate, 2012, p. 29). These projects intend to foster autonomous learning by helping the learners to increase their familiarity, comfort levels and abilities in performing self-assessments (Kodate & Foale, 2012). To do so, learners engage in a learning cycle which starts with collaboration with the teacher to develop unique can do statements, continues with the development of study materials and techniques to achieve the tasks implicated by their own can do statements and ends with an evaluation or reflection on the entire process (see Kodate, 2012, and Kodate & Foale, 2012, for further information about the type of materials used and the activities undertaken in this genre of course). Learners in these classes also all make frequent use of a Self-Access Center (SAC) designed to support the development of learning and language skills. In the SAC, they receive support from their language teachers and learning advisors (staff members dedicated to supporting students' learning), as well as materials intended to increase student self-directedness and improve their language abilities (see Thompson & Atkinson, 2010, for further description of the SAC). In spite of this training, participants had no known explicit familiarity with the CEFR or the CEFR-J.

Even though 57 participants does not represent a large sample size, it was not considered appropriate to include data from other cohorts of students (either internal or external to the university) as they would have experienced vastly different learning experiences to those in the current study. In the case of external students, their English curriculum may not have been CEFR-informed, or they may not have received training in self-assessment in the same way the participants had. In the case of including other internal students, there is no requirement to take the TOEIC, and it was thought that non-English majors could not be grouped with the English major students due to their lack of training in self-assessment and differences in the total time spent studying English. Ultimately, it was thought that the included participants represented a relatively homogenous group in terms of recent learning experiences and although a sample of this size may somewhat limit the generalizability of results, the more specific description of the learners and their learning experiences above intends to better allow readers to estimate and consider how the findings might be applicable to their own context.

Instruments

Survey

The can do statement survey was created and administered on www.surveymonkey.com©. Participants responded to the following stem-completion item: 'To what extent do you agree with the following statements (where 1 is strongly disagree and 4 is strongly agree)?'. A 4-point scale was employed to measure participants' self-assessed abilities on all 36 randomly ordered CEFR-J can do statements from the skills of listening and reading through the levels of A1.1 to B2.2 (covering a total of nine levels). A four-point scale was selected to reduce the Japanese tendency of selecting a neutral response (Dornyei & Taguchi, 2010). The can do statement survey was administered following completion of an academic year of full-time English study in the students' regular classroom and at the end of their usual class time. Depending on their year of study, the participants had completed one, two or three full years of their degree program.

All CEFR-J can do statements are available in both English and Japanese for free download at http://www.cefr-j.org/english/index-e.html although learners in the present study responded only to the Japanese descriptors as self-assessment has generally been found to be more accurate when administered in the native language (Oscarson, 1997). The English version of all of the can do statements used in the survey are shown in Table 3 (Appendix).

The TOEIC

The TOEIC listening and reading test is an English language proficiency test that measures the English-language reading and listening skills required in international workplaces. Each comprehension section of the test is graded, and then the scores combined to give the test-taker a score out of 990 (Educational Testing Service, hereafter ETS, 2015). Regarding the reading and listening TOEIC scores, the participants, being English majors, take the TOEIC annually and the closest test result to the time of survey was used. In this case, the TOEIC was taken within five days after the can-do statement survey.

Analysis

To address the research questions, the following analyses were performed. For the first research question, in determining the results of the self-assessments, the mean ratings for listening and reading can do statements from levels A1.1 to B2.2 from the survey were calculated for each participant. In order to explore the second research question of whether the learners' perceptions of difficulty were as predicted by the CEFR-J's difficulty hierarchy, a one-way ANOVA followed by LSD post-hoc analyses was performed to check for significant differences in difficulty between CEFR-J A and B levels to ensure that the B level can do statements were indeed receiving higher difficulty ratings. Such a check has previously been performed with non-English majors, and it was found that the higher level can do statements were generally associated with higher difficulty ratings (Runnels, 2013b, 2013c). For the remaining two questions, TOEIC scores were correlated with the self-assessment scores for the individual skills of listening and reading.

RESULTS

To address the first research question, the average difficulty ratings for first-year English majors' self-assessments on CEFR-J can do statements from level A1.1 to B2.2 using a four-point scale were calculated. For listening and reading respectively, a mean rating of 2.76 (*SD* = 0.84) and 2.77 (*SD* = 0.81) was found. Figure 1 shows the self-assessment ratings for listening and reading across all of the CEFR-J levels, where a general decrease in ratings is evident as the levels increase.

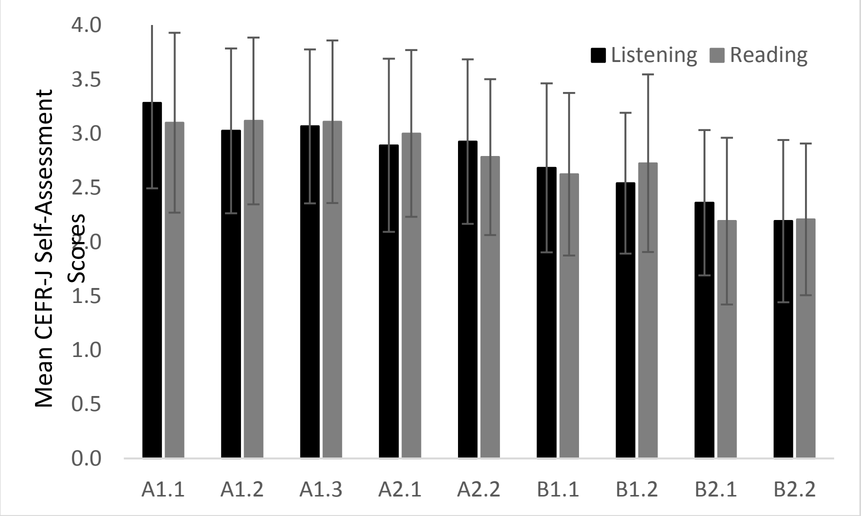

Figure 1. Self-assessments by English majors on the CEFR-J can do statements from levels A1 to B2 (Y-error bars show the standard deviation)

In terms of whether higher-level can do statements were rated as more difficult than lower-level statements, significant differences between difficulty ratings across all CEFR-J levels were found for both listening (*F*(8,504) = 20.27, *p* = .000) and reading (*F*(8,504) = 21.13, *p* = .000). Further testing also revealed that the B level can do statements for listening were rated as significantly more difficult (*M* = 2.47, *SD* = .42) than the A level statements (*M* = 3.10, *SD* =.59, *F*(1,112) = 43.36, *p* = .000). The same occurred for reading, with A level statements (*M* = 3.11, *SD* = .58) rated as easier than B level (*M* = 2.48, *SD* = 0.48, *F*(1,112) = 38.88, *p* = .000), which indicates that the can do statements were generally performing as predicted by the difficulty hierarchy of the CEFR-J. However, similar to what was found with non-English majors in Runnels (2013b, 2013c), there were no significant differences in ratings between most adjacent CEFR-J levels. For listening, the post hoc analyses revealed significant differences at an alpha level of .05 between A1.1 and A1.2 (*p* = .043), A2.2 and B1.1 (*p* = .043) and B2.1 and B2.2 (*p* = .036). The remaining five adjacent pairs exhibited no significant differences. For reading, significant differences were only found between B2.1 and B2.2 (*p* = .000) meaning that the remaining seven adjacent levels (from A1.1 through to B1.2) were not rated significantly differently when compared to the next closest CEFR-J level.

TOEIC scores ranged from 225 to 705 with an average score of 418 across all 57 participants. According to ETS (2008, 2013a), this range is equivalent to CEFR levels A2 to B1 (a TOEIC score 550 is required for a B1 level and 785 for B2). A mean of 418 is associated with a CEFR level of A2, with 550 required to reach a B1 level (ETS, 2008). In terms of the overall correlations, self-assessment scores exhibited no relationship with test scores, although this correlation was found to be not significant (*r*(511) = -.02, *p* = .580). An absence of significance was not found for each of the individual skills however, although only weak correlations were found for both listening (*r*(511) = .23, *p* = .000) and reading (*r*(511) = -.14, *p* = .001). However, removing the four most difficult levels, leaving just the A level can do statements to determine if descriptors of higher difficulty were affecting self-assessment scores, altered the correlations significantly. The correlation between the A level self-assessment ratings and TOEIC scores for listening was significant and moderate (*r*(169) = .33, *p* = .000), although for reading, it was found to be weakly negative (*r*(169) = -.16, *p* = .042). Examining the differences in correlations between the levels even further, for the A1 level can do statements the r-value increased to *r*(55) = .44, *p* = .001 for listening while the reading correlation was insignificant and relatively weak (r(55) = -.22, *p* = .097). Within the A2 level, the correlations for listening and reading respectively were *r*(55) = .29 (*p* = .002), and *r*(55) = -.12 (*p* = .193). This raises the possibility that the correlations differ according to CEFR-J level, and these tests were therefore repeated accordingly. The same tests were also performed within the B1 and the B2 statements. The results are shown in Table 2.

Table 2

The Correlations Found Between Self-assessment and TOEIC Scores for the Skills of Reading and Listening at Each CEFR-J Level

CEFR-J level	Correlation between TOEIC and CEFR-J self-assessment scores[*]	
	Listening	Reading
A1.1	*r* = 0.44, *p* = .001	*r* = -0.22, *p* = .097
A1.2	*r* = 0.31, *p* = .021	*r* = -0.19, *p* = .164
A1.3	*r* = 0.31, *p* = .019	*r* = -0.19, *p* = .148
A2.1	*r* = 0.24, *p* = .010	*r* = -0.14, *p* = .313
A2.2	*r* = 0.24, *p* = .078	*r* = -0.11, *p* = .399
B1.1	*r* = 0.16, *p* = .244	*r* = -0.21, *p* = .111
B1.2	*r* = 0.28, *p* = .036	*r* = -0.16, *p* = .222
B2.1	*r* = 0.11, *p* = .414	*r* = -0.27, *p* = .045
B2.2	*r* = 0.13, *p* = .329	*r* = 0.02, *p* = .873

Note. [*]Each r-value has 55 degrees of freedom.

As can be seen in Table 2, for listening, there is a general increase in r-values as the CEFR-J level decreases, from a moderate strength correlation at the A1.1 level, to a weak correlation at the B2.2 level, although the correlations at the higher end of the difficulty spectrum are not significant. This is contrary to the reading correlations, which are mostly weak, not significant, and do not vary as widely across the CEFR-J levels. An unexpected result is that the correlations, albeit not significant, are negative, which means that as the participants' TOEIC scores increased, so did their difficulty ratings. In other words, the reading correlations are a lot less reliable than the listening correlations, but may be more consistent.

DISCUSSION

English majors from a Japanese university rated CEFR-J listening and reading can do statements for difficulty, indicating that, on average, they tended to agree that they could perform the communicative tasks implicated by all of the statements from levels A1.1 to B2.2. As was shown in Figure 1, their agreement generally increased as the CEFR-J level decreased, meaning that they found the higher level can do statements to be less easy to perform. These results were also confirmed when tests for significant differences were conducted between the varying CEFR-J levels, whereby differences were found both overall and between the A and B levels for each skill. However, when each CEFR-J sub-level was tested for differences between it and its adjacent levels, only one significant difference was revealed for reading and listening (between levels B2.1 and B2.2). This lack of significant differences between adjacent CEFR-J levels is similar to previous findings with non-English majors (Runnels, 2013b; 2013c) and suggests that perhaps the division of the A1 level into three sub-levels of (A1.1, A1.2 and A1.3), A2 into two (A2.1 and A2.2) is too many for CEFR-J users to consistently distinguish between. Nevertheless, the CEFR-J's can do statements as a self-assessment instrument performed generally as expected in that the difficulty ratings increased in accordance with their CEFR-J level.

In terms of the relationship between TOEIC scores and CEFR-J self-assessment ratings, only moderate correlations for listening and weak correlations for reading were found, the latter's of which were all not significant (but one). Contrarily, the listening correlations were mostly significant. The findings in the current study suggest that language proficiency as measured by the TOEIC does not reliably correlate with self-assessment scores by Japanese English language learners on CEFR-J can do statements although this unreliability is stronger for reading than for listening. According to Ross (1998) and Spolsky (1992) who observed that self-assessment accuracy correlates with experience with the language skill, the current finding may suggest that learners have significantly more experience with listening than with reading in English. Greater learner experience with a skill is certainly a possible explanation for the stronger correlation in listening, but is unlikely in the current case: Even though students' make near daily usage of a SAC where they interact with peers, teachers and learning advisors in English only (and are therefore hearing a lot of English), the students are also reading in English on a daily basis in their English classes and also in the SAC. Ross (1998) found that of all four language skills, reading self-assessment correlated most reliably with reading test scores, since "exposure to the written word predates extensive opportunities for listening… practice" (p. 6, as cited in Ito et al., 2005). Although this finding was not corroborated in the present investigation, experience with the language skill may remain a major variable in influencing self-assessment accuracy.

It is also likely that self-assessment accuracy was affected by task difficulty, in that ratings on easy tasks received far more consistent or accurate ratings than difficult tasks where the learner may struggle to imagine successfully performing the task. In fact, not only is this supported by the increase in strength of the correlations that was found as the CEFR-J levels decreased for listening, Sundstroem (2005) has also previously found that the reliability of can do statement self-assessment is affected by task difficulty. As North (2007) explains, the notion of can do statement mastery is abstract, and when no specific definition for mastery exists, this will naturally lead to differing perceptions of difficulty of the task across learners, which increase with task complexity.

What remains unclear though, is whether the perceived increase in task difficulty (which seems to be associated with lower accuracy in self-assessment ratings) is indeed a result of a greater task difficulty, or whether it is due to a lack of familiarity with the task, not being able to imagine performing the task or whether it is lower confidence in selecting a rating indicating higher mastery. The latter may certainly be possible given the negative correlations for reading whereby participants with greater reading proficiency (as measured by their TOEIC reading scores), indicated they were less able at performing the reading tasks implicated by the can do statements. In fact, the reading correlations were all around the r = -.20 mark, which may be accounted for by a consistent low confidence in participants' own reading abilities. In any case, although experience with language skill has been associated with increased self-assessment accuracy (Ross, 1998; Spolsky, 1992), likewise have task difficulty (Sundstroem, 2005) and confidence (MacIntyre et al., 1997) been noted as influential factors. The results also suggest that perceived task difficulty may be mediated by familiarity with the task or self-confidence. Each of these factors certainly appear to be relevant to self-assessment accuracy and should be considered in more detail in further study.

The lack of strong or moderate correlations between self-assessment and test scores could also be due to the fact that the participants did not understand the can do statements (Runnels, 2014a; Spolsky, 1992). Numerous studies suggest that misunderstanding of the statements is unlikely in this case though, as they were developed and tested extensively by Japanese university students (Negishi, 2011; Negishi et al., 2013; Tono & Negishi, 2012), they have been previously rated reliably by Japanese university students for difficulty (Runnels, 2013b, 2013c), and were presented to the learners in Japanese (Oscarson, 1997). Another possibility is that the self-assessment training undertaken by students was insufficient, since "self-assessment depends on a complex of skills" (Little, 2005, p. 332). This is unlikely though, as participants had undergone extensive training in performing self-assessment, and lack of correlations with test scores and inconsistent ratings across skill were nonetheless found.

Alternatively, perhaps the weak correlation overall for reading and only moderate correlation for listening is due to Blanche and Merino's (1989) suggestion that self-assessment accuracy depends on the similarity of the content of the instruments involved. Ross (1998, p. 16, as cited in Ito et al., 2005) concurs: Variation in self-assessment accuracy is determinate of whether "the criterion variable is one that exemplifies achievement of functional 'can do' skills on the self-assessment battery". In other words, a mismatch in content between the test and self-assessment material will result in a lower level of self-assessment accuracy. This is a possible limitation to the current study where the test's focus is largely centered on the language of commerce, and not on general English involved in everyday situations (the focus of the participants' curriculum and the self-assessment battery). In order to address Edele et al.'s (2015) complaint that the quality of instruments used in previous studies of self-assessment accuracy was questionable, two widely known and used instruments - the TOEIC and (a modified version of) the CEFR - were employed in the current study. Although the suggestion to use instruments of established quality is a valid one, the choice of instruments in the current study appears to be misguided and may account for the lack of correlations. A self-assessment using a TOEIC-informed battery of can do statements which reflect the everyday skills of people working in an international environment, or a CEFR-informed test, such as the Oxford Online Placement Test (Pollit, 2009), or system such as DIALANG, which provides estimates of CEFR level based on performance on tasks derived from the CEFR's can do statements (Alderson & Huhta, 2005), may be better suited to address Edele et al.'s (2015) critique. In employing such instruments, the comparison of self-assessment to proficiency would still be derived from widely known and used instruments, but the content would be arguably more well-matched.

Nonetheless, the issues of mismatch in content between self-assessment battery and testing instrument returns to Ito et al.'s (2005, p. 2) comments about the difficulty of interpreting norm-referenced test scores, including TOEIC, since they do not give "concrete ideas of what tasks a person in a certain score band is able to do with English". ETS (2008) has attempted to address these criticisms by providing can do lists for score bands for each skill of listening and reading, but care should nevertheless be taken not to apply TOEIC scores to a context with mismatched content (Ito et al., 2005). For instance, if a student needs a TOEIC score to graduate, the language program should perhaps make clear that it is geared towards preparing the students to score well on the test, at the possible expense of helping their students become more proficient language users. Institutions may also find that other exams such as the TOEFL® (Test of English as a Foreign Language, Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000; McNamara, 2001), IELTS™ (International English Language Testing System, ielts.org, n.d.), or those designed in-house may provide a valid and more suitable means for an institution to measure proficiency (and self-assessment accuracy).

These findings have implications for how the CEFR-J's (and by extension, the CEFR's) can do statements can be used as a self-assessment instrument by Japanese language learners. As Gardner (2000) has shown, if the intention of the assessment is to motivate the learner, then one which allows the learner to obtain immediate and individualized feedback should be conducted so that reflection on their goals, strategies and ensuing achievement can occur. Conversely, if the intention is to allow for an evaluation of learning materials by the teacher or learner, then a different approach to self-assessment, or a different self-assessment instrument may be required. As Gardner (2000, p. 53) puts it, "the degree to which a certain unreliability can be tolerated depends on the uses to which the assessments will be put. Where used for individual monitoring of progress, absolute reliability may be of less importance [as] assessments which are not totally reliably still offer many. . .benefits". In the same way that the conduct of assessment is linked to the usage of assessment scores, likewise does self-assessment score usage determine the conduct of the self-assessment (or vice versa).

Although a limitation to this study is that participants' understanding and their perceptions of the difficulty of the implicated tasks of the can do statements is not known, it seems clear that self-assessment, even for highly-trained Japanese learners of English, is subject to a number of mediating factors. Future research involving interviews with participants may shed more light into the range of factors which impacted the processes and results of language learners' self-assessments, whether those be related to sample size, age of participants, socio-economic background, format of tests, whether situated in a target language environment and, what was being compared to the self-assessment scores (Gardner, 2000), or, as was suggested by the current study: the task difficulty, content overlap between instruments, participants' familiarity with task, experience with language skill, general self-confidence in their own abilities and the aims of the self-assessment and intended usage of the self-assessment and test scores.

However, it seems that despite the lack of conclusiveness regarding self-assessment accuracy (in both the current and previous studies), many researchers maintain "a belief in the value of self-assessment" (Gardner, 2000, p. 53). For instance, Dickinson (1987) believes that self-assessment is an important skill for all language learners to acquire, while Janssen-van Dieten (1989) finds that its value is in the positive influence it effects onto the learning process, rather than its accuracy. Additionally, Thomson (1996) and Gardner (2000) list many positive effects of training learners in self-assessment, despite its unreliability. In addition, learners from the present study indicated that can do statements aided in the achievement of their learning goals, were effective "as tools for exerting control over self-evaluation and assessment, . . .a useful means by which to identify their strengths and weaknesses, . . .[and] helped them understand what and how they learned" (Kodate & Foale, 2012, pp. 37-38).

CONCLUSION

The current study suggested that Japanese institutions, particularly those in the tertiary education sector, be cautious about requiring or using TOEIC scores as a measure of language proficiency, especially if the content of the TOEIC is not very closely matched with the pedagogical content of the language program. It was also noted that CEFR-J (or CEFR) self-assessments may not be able to be validly employed to aid with the interpretation of TOEIC scores or any other test that is not CEFR-informed. This does not mean that the CEFR-J cannot be used as an effective self-assessment tool on an individual learner basis, or within a context where a language curriculum or training program does emphasise the teaching and learning of language for use in everyday situations.

In Japan, despite increased educational spending in the area of English language education (Jones, 2011; Ministry of Finance Japan, 2003), and continually high numbers of Japanese TOEIC examinees (ETS, 2013b; McCrostie, 2010), there is currently no consistently used system for dictating the content of language instruction or measuring proficiency or progress of Japanese English language learners (Negishi et al., 2013; Runnels, 2014b). Even though the CEFR-J has been introduced to address this, its current form consists solely of a modified set of CEFR levels and can do statements and how these should be or has been used, particularly for learner self-assessment and the development of autonomous learners remains to be seen. Self-assessment is nonetheless an integral part of a CEFR-informed approach to language teaching, learning and assessment, and because of this, further study on the CEFR-J and self-assessment by Japanese language learners is required. Doing so would also aid in the interpretability of test scores, and increase understanding about self-assessment accuracy, which would in turn help teachers foster the development of autonomy in their learners, or help a learner better understand their estimates of their own proficiency.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*, 659-*663*.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*, 301-320.

Bachman, L., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*, 14-29.

Blanche, P. (1990). Using standardised achievement and oral proficiency tests for self-assessment purposes: The DLIFC study. *Language Testing*, *7*, 202-229.

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

Blue, G. M. (1988). Self-assessment: The limit of learner independence. In A. Brookes & P. Grundy (Eds.), *Individualisation and autonomy in language learning* (pp. 100-118). London: Modern English Publications in association with the British Council (Macmillan).

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System*, *40*, 144-160.

Chapman, M. (2003). TOEIC®: Tried but undertested. *Shiken Research Bulletin*, *7*(3), 2-7.

Chen, Y. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, *12*, 235-262.

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 66-75). Tokyo, Japan: Japan Association of Language Teachers.

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

Council of Europe. (2005). *Survey on the use of the Common European Framework of Reference for Languages (CEFR): Synthesis of results*. Strasbourg, France: Council of Europe.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing*, *2*, 164-179.

Dickinson, L. (1987). *Self-instruction in language learning*. Cambridge: Cambridge University Press.

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

Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research*, *52*, 99-123.

Educational Testing Service. (2008). *Correlation table TOEIC® listening and reading scores descriptors and European CEFR levels*. Retrieved from http://www.toeic.ca/fileadmin/free_resources/ETS_Global_master/TOEIC_L_R_can-do_table.pdf

Educational Testing Service. (2013a). *Mapping the TOEIC and TOEIC Bridge™ tests on the Common European Framework Reference (CEFR)*. Retrieved from http://www.ets.org/s/toeic/pdf/toeic_cef_mapping_flyer.pdf

Educational Testing Service. (2013b). *Report on test takers worldwide: The TOEIC® listening and reading test*. Retrieved from https://www.ets.org/s/toeic/pdf/ww_data_report_unlweb.pdf

Educational Testing Service. (2015). *The TOEIC*. Retrieved from https://www.ets.org/toeic

Eiken. (n.d.). *What is BULATS?* Retrieved from http://www.eiken.or.jp/bulats/en/

Finnie, R., & Meng, R. (2005). Literacy and labour market income: Self-assessment versus test score measures. *Applied Economics*, *37*, 1935-1951.

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. In A. Psyaltou-Joycey & M. Matthaioudakis (Eds.), *Advances in Research on Language Acquisition and Teaching: Selected Papers* (pp. 15-26). Thessaloniki, Greece: GALA.

Gardner, D. (2000). Self-assessment for autonomous language learners. *Links & Letters*, *7*, 49-60.

Gardner, D., & Miller, L. (1999). *Establishing self-access: From theory to practice*. Cambridge: Cambridge University Press.

Gilfert, S. (1996). A review of TOEIC. *The Internet TESL Journal*, *2*(8). Retrieved from http://iteslj.org/Articles/Gilfert-TOEIC.html

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

Green, A. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge: Cambridge University Press.

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing*, *7*, 174-201.

Holec, H. (1981). *Autonomy and foreign language learning*. Oxford: Pergamon.

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

ielts.org. (n.d.). *What is IELTS?* Retrieved from http://www.ielts.org/test_takers_information/what_is_ielts.aspx

Ikeno, O. (2002). *The Japanese mind: Understanding contemporary culture*. North Clarendon, VT: Tuttle.

Ito, T., Kawaguchi, K., & Ohta, R. (2005). *A study of the relationship between TOEIC* *scores and functional job performance: Self-assessment of Foreign Language Proficiency* (TOEIC Research Report 1). Tokyo, Japan: Institute for International Business Communication. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/newsletter/TaeIto_E.pdf

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P. B., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RM-00-03.pdf

Janssen-van Dieten, A. (1989). The development of a test of Dutch as a foreign language: The validity of self-assessment by inexperienced subjects. *Language Testing, 6*, 30-46.

Jones, R. S. (2011). Education Reform in Japan. *OECD Economics Department Working Papers, 888.* Retrieved from http://dx.doi.org/10.1787/5kg58z7g95np-en

Kodate, A. (2012). The JASAL Forum 2011: Growing trends in self-access learning. *Studies in Self-Access Learning Journal*, *3*, 122-132.

Kodate, A., & Foale, C. (2012). The effectiveness of 'can-do' statements as tools to enhance autonomous language learning skills. *Hiroshima Bunkyo Women's University Journal*, *47*, 31-40. Retrieved from http://harp.lib.hiroshima-u.ac.jp/h-bunkyo/metadata/12099

Kubota, R. (2011). Questioning linguistic instrumentalism: English, neoliberalism, and language tests in Japan. *Linguistics and Education, 22*, 248-260.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, *19*, 673-687.

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

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

Little, D. (2007). The Common European Framework of Reference for languages: Perspectives on the making of supranational language education policies. *The Modern Language Journal*, *91*, 645-655.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*, 265-287.

Martyniuk, W., & Noijons, J. (2007, February). *Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States*. Paper presented at the Council of Europe Intergovernmental Language Policy Forum, Strasbourg, France. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Survey_CEFR_2007_EN.doc

Matsuno, S. (2000). Self-, peer-, and teacher-assessments in Japanese university EFL writing. *Language Testing*, *26*, 75-100.

McCrostie, J. (2010). The TOEIC in Japan: A scandal made in heaven. *Shiken Research Bulletin*, *14*(1), 2-10. Retrieved from http://jalt.org/test/mcc_1.htm

McNamara, T. (2001). The challenge of speaking: Research on the testing of speaking for the TOEFL. *Shiken Research Bulletin*, *5*(1), 2-3. Retrieved from http://jalt.org/test/mcn_1.htm

MEXT. (2013). 各中・高等学校の外国語教育における「CAN DO リスト」の形での 学習到達目標設定のための手引き [*Guides for setting learning goals in the form of CAN DO lists in foreign language education in middle and high school*]. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/__icsFiles/afieldfile/2013/05/08/1332306_4.pdf

Ministry of Finance Japan. (2003). *Understanding the Japanese budget*. Retrieved from http://www.mof.go.jp/english/budget/budget/fy2003/brief/2003-10.htm

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

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. (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 Kraków Conference* (pp. 135-163). Cambridge: Cambridge University Press.

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, France: Council of Europe.

North, B. (2007, February). *The CEFR Common Reference Levels: Validated reference points and local strategies*. Paper presented at The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities, Strasbourg. Retrieved from http://www.coe.int/T/DG4/.../North-Forum-paper_EN.doc

North, B., Ortega, A., & Sheehan, S. (2010). *A core inventory for general English, British Council/EAQUALS*. Retrieved 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*, 217–262.

O'Dwyer, F., & Runnels, J. (2014). Bringing learner self-regulation practices forward. *Studies in Self-Access Learning Journal*, *5*, 404-422.

O'Dwyer, F., Imig, A., & Nagai, N. (2014). Connectedness through a strong form of TBLT, classroom implementation of the CEFR, cyclical learning, and learning-oriented assessment. *Language Learning in Higher Education*, *3*, 231-253.

O'Dwyer, F., Nagai, N., Imig, A., Naganuma, N., Schmidt, G., & Hunke, M. (n.d.). *Critical, constructive assessment of CEFR-based language teaching in Japan and beyond*. Retrieved from https://sites.google.com/site/flpsig/critical-constructive-assessment-of-cefr

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. *Language Testing and Assessment*, *7*, 175-187.

Pierce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion and locus of control. *Applied Linguistics*, *14*, 25-42.

Pollit, A. (2009). *The Oxford Online Placement Test: The meaning of OOPT scores*. Retrieved from http://www.oxfordenglishtesting.com/uploadedFiles/Buy_tests /oopt_meaning.pdf

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

Runnels, J. (2013a). A preliminary exploration of the relationship between student ability, self-assessment and teacher assessment on the CEFR-J's can do statements. *Framework & Language Portfolio Newsletter*, *9*, 6-18.

Runnels, J. (2013b). Preliminary Validation of A1 and A2 sub-levels of the CEFR-J. *Shiken Research Bulletin*, *17*, 3-10.

Runnels, J. (2013c). Examining the difficulty pathways of can-do statements from a localized version of the CEFR. *Applied Research on the English Language*, *2*(1), 25-32.

Runnels, J. (2014a). An exploratory reliability and content analysis of the CEFR-Japan's A-level can-do statements. *JALT Journal*, *36*, 69-89.

Runnels, J. (2014b). The CEFR-J: The story so far. *Framework & Language Portfolio SIG Newsletter*, *12*, 9-20.

Schmidt, M.S., Naganuma, N., O'Dwyer, F., Imig, A., & Sakai, K. (Eds.). (2010). *Can do statements in language education in Japan and beyond*. Tokyo, Japan: Asahi Press.

Shibata, J., & Inoue, H. (2005). *A development of a context-based curriculum for TOEIC Level D students of Kosen (National College of Technology)* (TOEIC Research Report 2). Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/ newsletter/JunkoShibata_E.pdf

Shimo, E., & Nitta, K. (2011). Developing can-do lists as a self-evaluation tool for university-level English classes. *Kinki University Liberal Arts and Foreign Language Education Center Bulletin*, *2*(1), 225-245. Retrieved from http://jairo.nii.ac.jp/0066 /00007627/en

Spolsky, B. (1992). Diagnostic testing revisited. In E. Shohamy & R. A. Walton (Eds.), *Language assessment and feedback: Testing and other strategies* (pp. 29-39). Dubuque, IA: Kendall/Hunt.

Sundstroem, A. (2005). Self-assessment of knowledge and abilities: A literature study. *EM*, *54*, 1-36. Retrieved from http://www.edusci.umu.se/digitalAssets/60 /60577_em541.pdf

Takada, N., & Lampkin, R. (1996). *The Japanese way: Aspects of behavior, attitudes and customs of the Japanese*. New York: McGraw-Hill.

Takada, T. (2014, May). *Contextualization of the CEFR in English language teaching in Japan*. Paper presented at Critical, constructive assessment of CEFR-based language teaching in Japan and beyond, Nagoya, Japan.

Tavakoli, M., & Ghoorchaei, B. (2009). On the relationship between risk-taking and self-assessment of speaking ability: A case of freshman EFL learners. *The Journal of Asian TEFL*, *6*(1), 1-27. Retrieved from www.asiatefl.org/main/download_pdf.php?i

=235&c=1419310604

Thompson, G., & Atkinson, L. (2010). Integrating self-access into the curriculum: Our experience. *Studies in Self-Access Learning Journal*, *1*, 47-58. Retrieved from http://sisaljournal.org/archives/jun10/thompson_atkinson/

Thomson, C. K. (1996). Self-assessment in self-directed learning: Issues of learner diversity. In R. Pemberton, E. Li, W. Or, & H. Pierson (Eds.), *Taking control: Autonomy in language learning* (pp. 77-91). Hong Kong: Hong Kong University Press.

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. (2007). *Modern languages in the Council of Europe 1954-1997: International co-operation in support of lifelong language learning for effective communication, mutual cultural enrichment and democratic citizenship in Europe*. Strasbourg, France: Council of Europe.

University of Cambridge Local Examinations Syndicate. (2016). *BULATS*. Retrieved from http://www.bulats.org/

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

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

Wilson, K. M. (1989). *Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC testing context* (TOEIC Research Report 1). Princeton, NJ: ETS. Retrieved from https://www.ets.org/research/policy_research_reports/publications /report/1989 /hwwb

Woodford, P. E. (1982). *An introduction to TOEIC: The initial validation study* (TOEIC Research Summaries No. 0). Princeton, NJ: ETS.

ACKNOWLEDGMENT

The author would like to acknowledge the significant contributions of Jack Bower of Hiroshima Bunkyo Women's University, where the study was conducted.

CORRESPONDENCE

Judith Runnels, Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, Luton, U.K.
E-mail address: judith.runnels@study.beds.ac.uk

PUBLISHING RECORD

Manuscript received: September 11, 2014; Revision received: July 16, 2015; Manuscript accepted: August 6, 2015

APPENDIX

Table 3. The English Versions of CEFR-J A1.1-B2.2 Listening and Reading Can-do Statements Administered to English Majors

CEFR-J Level	Listening	Reading
A1.1	I can understand short, simple instructions such as "Stand up." "Sit down." "Stop." etc., provided they are delivered face-to face, slowly and clearly. I can catch key information necessary for everyday life such as numbers, prices, dates, days of the week, provided they are delivered slowly and clearly.	I can understand a fast-food restaurant menu that has pictures or photos, and choose the food and drink in the menu. I can read and understand very short, simple, directions used in everyday life such as "No parking", "No food or drink", etc.
A1.2	I can understand short conversations about familiar topics (e.g., hobbies, sports, club activities), provided they are delivered in slow and clear speech. I can catch concrete information (e.g., places and times) on familiar topics encountered in everyday life, provided it is delivered in slow and clear speech.	I can understand very short reports of recent events such as text messages from friends' or relatives', describing travel memories, etc. I can understand very short, simple, everyday texts, such as simple posters and invitation cards.
A1.3	I can understand instructions and explanations necessary for simple transactions (e.g., shopping and eating out), provided they are delivered slowly and clearly. I can understand phrases and expressions related to matters of immediate relevance to me or my family, school, neighborhood etc., provided they are delivered slowly and clearly.	I can understand short narratives with illustrations and pictures written in simple words. I can understand texts of personal interest (e.g., articles about sports, music, travel, etc.) written with simple words supported by illustrations and pictures.

CEFR-J Level	Listening	Reading
A2.1	I can understand short, simple announcements (e.g., on public transport or in stations or airports) provided they are delivered slowly and clearly. I can understand the main points of straightforward factual messages (e.g., a school assignment, a travel itinerary), provided speech is clearly articulated in a familiar accent.	I can find the information I need, from practical, concrete, predictable texts (e.g., travel guidebooks, recipes), provided they are written in simple English. I can understand explanatory texts describing people, places, everyday life, and culture, etc., written in simple words.
A2.2	I can understand and follow a series of instructions for sports, cooking, etc. provided they are delivered slowly and clearly. I can understand instructions about procedures (e.g., cooking, handicrafts), with visual aids, provided they are delivered in slow and clear speech involving rephrasing and repetition.	I can understand short narratives and biographies written in simple words. I can understand the main points of texts dealing with everyday topics (e.g., life, hobbies, sports) and obtain the information I need.
B1.1	I can understand the gist of explanations of cultural practices and customs that are unfamiliar to me, provided they are delivered in slow and clear speech involving rephrasing and repetition. I can understand the main points of extended discussions around me, provided speech is clearly articulated and in a familiar accent.	I can understand the main points of extended discussions around me, provided speech is clearly articulated and in a familiar accent. I can understand the main points of English newspaper and magazine articles adapted for Educational purposes.
B1.2	I can understand the majority of the concrete information content of recorded or broadcast audio material on topics of personal interest spoken at normal speed.	I can understand the main points of short radio news items about familiar topics if they are delivered in a clear, familiar accent.

CEFR-J Level	Listening	Reading
	I can understand the main points of short radio news items about familiar topics if they are delivered in a clear, familiar accent.	I can search the internet or reference books, and obtain school- or work-related information, paying attention to its structure and given the occasional use of a dictionary, I can understand it, relating it to any accompanying figures or tables.
B2.1	I can understand the main points of a conversation between native speakers in television programmes and in films, provided they are delivered at normal speed and in standard English. I can follow extended speech and complex lines of argument provided the topic is reasonably familiar.	I can follow extended speech and complex lines of argument provided the topic is reasonably familiar. I can read texts dealing with topics of general interest, such as current affairs, without consulting a dictionary, and can compare differences and similarities between multiple points of view.
B2.2	I can follow a variety of conversations between native speakers, in television programmes and in films, which make no linguistic adjustments for non-native speakers. I can understand the speaker's point of view about topics of current common interest and in specialised fields, provided it is delivered at a natural speed and articulated in standard English.	I can understand the speaker's point of view about topics of current common interest and in specialised fields, provided it is delivered at a natural speed and articulated in standard English. I can scan through rather complex texts e.g. articles and reports, and can identify key passages. I can adapt my reading speed and style, and read accurately, when I decide closer study is worthwhile.

日本的英語學習者使用 CEFR-J 的自我評量及多益英語測驗分數相關性研究

Judith Runnels
英國貝德福德大學

多益英語測驗 TOEIC®（國際職場英語溝通測驗），自 1979 年發行以來持續為日本教育機構及各大公司廣泛地使用。但是該測驗也一再被質疑，其測驗結果未能有效說明受試者的語言能力。為避免片面強調多益測驗成績高低，也為了能更有效地解讀多益測驗成績。相關單位也開始使用其他的語言能力評量方式。其中一項顯著的轉變就是採用 CEFR（歐洲語言共同參考架構）。「歐洲語言共同參考架構」係以學習者為中心，針對語言的教與學、以及評量所提出的一套參考標準。CEFR 不但可以提升學習者的自主學習，也能透過「能力指標說明」（can do statements）的使用，來幫助學習者完成自我評量。「能力指標說明」描述語言學習者在特定場合的語言溝通能力。由於日本愈來愈重視 CEFR 在語言學習上的評量功能，因此探究語言能力以及自我評量之間關連性的研究需求也隨之增加。本研究旨在探討以日文為母語的英語學習人士的聽讀自我評量分數，以及多益測驗成績之間的關連性為何。該自我評量表為 CEFR-J（Common European Framework of Reference-Japan）乃歐洲語言共同參考架構 CEFR 的修改版。研究結果發現自我評量分數與多益測驗成績略相關，但在閱讀方面並無相關。文中並探討影響下列事項的可能因素：學習者的自我評量意願、自我評量系統用於日本的英語學習者的功效，以及 TOIEC® 分數的解讀。

關鍵詞： 多益英語測驗、歐洲語言共同參考架構、CEFR-J、自我評量、語言能力、能力指標說明