The Asian Journal of Applied Linguistics
Vol. 1 No. 2, 2014, pp. 167-77

AJAL

Japanese English learners' self-assessments on the CEFR-J's A-level can-do statements using four and five-point response scales

Judith Runnels
University of Bedfordshire, UK

Both the CEFR (Common European Framework of Reference) and the CEFR-J (CEFR-Japan) use illustrative can-do descriptors to describe a learner's communicative competences in five language skills across six levels of language proficiency. This paper reports on Japanese English learners' self-assessments on the CEFR-J's 50 A-level descriptors using either a four-point or a five-point scale to determine if a neutral response option (*neither agree nor disagree*) influenced participants' responses. Self-assessment by Japanese language learners has been shown to be subject to cultural factors related to social desirability phenomena, resulting in high selection rates of mid-scale response options no matter the content of the item or the size of the scale. Overall, no significant differences between mean responses on a four-point (no neutral category) and a five-point (contains an inherent mid-point) rating scale were found following controls for scale size. Conversely, significant interactions were found for rating scale, skill (reading and spoken production) and descriptor difficulty level (A1.1 and A2.2). When the distance between responses and the scale mid-point was measured and compared across rating scales to determine whether the inclusion of a neutral option appeared to influence selection rates, no significant differences were found for 68% of all descriptors. While inclusion of a middle response option had far lesser impact on responses than has been previously shown, further research is required to determine the impact of differing scale types on Japanese English learners' self-assessments. This paper discusses the influence on responses from socio-cultural factors, response styles, task-familiarity, language skills, the number of response scale categories and language proficiency.

Keywords: Common European Framework of Reference; CEFR-Japan; can-do statements; self-assessment; scale size; Japan

Background

One of the strengths of the Common European Framework of Reference (CEFR) developed by the Council of Europe (2001) is that it allows educational institutions to compare the outcomes and content of language programmes (North, 2000) by providing a guide as to what a language learner can do with language at any given time (Council of Europe, 2001). The six main proficiency levels consist of scales of illustrative descriptors or can-do statements that describe second language learner proficiency across several language skills: listening, reading, spoken interaction, spoken production and writing (Council of Europe, 2001). In addition to usage at an institutional or curricular level, the CEFR can also "support the development of learner autonomy and learner self-assessment" (Little, 2006, p. 176). Glover (2011) has found that using CEFR can-do statements increases learners' self-awareness of language use, subsequently improving their overall development as a language user. Performing a self-assessment, typically by "using checklists based on the CEFR's common reference

ISSN 2308-6262
http://caes.hku.hk/ajal

levels" (Little, 2006, p. 176), often proceeds as follows: a learner reads a can-do statement such as the A1 Reading Comprehension descriptor (see below) and then decides whether she or he can perform the implicated task (Glover, 2011; Little, 2005). Doing this across a range of statements can then produce an estimation of ability across the CEFR's six proficiency levels (A1, A2, B1, B2, C1, C2, arranged in order of increasing difficulty) which is relevant to that learner.

> A1 Reading Comprehension descriptor:
> I can understand simple forms well enough to give basic personal details (e.g. name, address, date of birth) (Council of Europe, 2001)

Conversely, the CEFR is criticised for not being based on second-language acquisition theory or on performance samples from actual learners (Hulstijn, 2007; Westhoff, 2007). Furthermore, the scales by which proficiency is measured lack reification and do not provide any information regarding how tests can be developed or compared (Fulcher, 2003, 2010; Weir, 2005). Indeed, the underpinning of the hierarchy of the framework, which is used to measure proficiency or progress, is primarily based on teacher perceptions of what second language learner proficiency entails at different levels (Fulcher, 2004; North, 2007).

Despite its opponents, the CEFR has impacted second language education not only in Europe, but also in other regions of the world (Bärenfänger & Tschirner, 2008; Parmenter & Byram, 2010) including countries such as Argentina (Porto & Barboni, 2012) and Canada (Faez, Majhanovich, Taylor, Smith, & Crowley, 2011). Interest is also increasing in Asia (Wang, Kuo, Tsai, & Liao, 2012), where national tests of English language proficiency in Hong Kong and Taiwan have been mapped to the CEFR (Hsiao & Broeder, 2013; Wu, 2012). Yoneka (2011) has argued for a Common Asian Framework of Reference.

Context

Significant interest has been shown in implementing the CEFR in foreign language programmes of tertiary institutions of Japan (O'Dwyer & Nagai, 2011). Indeed, educators and researchers have already employed it in a number of ways at numerous tertiary institutions (see, for example: Horiguchi, Harada, Imoto, & Atobe, 2010; Kizman & Nitta, 2010; Nagai, 2010; Nakano, Tsutsui, & Kondo, 2010; O'Dwyer, 2013; Semmelroth, 2013). However, it has also been found that eighty percent of Japanese English learners are at a CEFR A-level of proficiency and that there are very few learners at a C-level (Negishi, 2012). It was thus decided that the CEFR levels in their current form were insufficient to characterise and differentiate between the span of Japanese learners of English. This resulted in the development of an alternate system intended to better meet the needs of Japanese English language learners and to address the lack of a consistently used system for the measurement of achievement of English language learners across Japanese institutions (Negishi, 2011, 2012). Released in March 2012 (TUFS Tonolab, 2012), known as the CEFR-J and developed courtesy of Grant-in-Aid research grants awarded to the Tokyo University of Foreign Studies (TUFS), it contains several modifications to the version from which it is derived (Negishi, Takada, & Tono, 2013). Changes to the original consist of adaptations of CEFR can-do statements to increase their suitability for a Japanese context, the sub-division of the A and B levels from the CEFR's original four levels (A1, A2, B1, B2) to nine (A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2), and the addition of a Pre-A1 level (Negishi et al., 2013).

Self-assessment

Despite interest in the CEFR-J's implementation, how the system can function as an assessment or a self-assessment instrument in Japan is under-researched, partly because the CEFR does not make any specific recommendations in terms of how can-do statements should be employed for such purposes. Previous work outside of a CEFR context has shown that self-assessment by Japanese learners is a complicated process, subject to uniquely Japanese cultural factors (Ikeno, 2002; Matsuno, 2009; Takada & Lampkin, 1996). For Japanese survey-takers the response scale that is utilized in the survey will significantly impact findings and subsequent conclusions (Pashupati, Courtright, & Pettit, 2013). In fact, Japanese survey-takers in general tend to select a neutral response as an option (Dörnyei & Taguchi, 2010; Pashupati et al., 2013; Ryan, 2009) in order to demonstrate greater modesty which is considered a virtuous trait in Japan (Ikeno, 2002; Matsuno, 2009; Takada & Lampkin, 1996). Conversely, even though Murata and Onodera (2011) found no "clear trend that particular response options were selected as a result of [the] existence/non-existence of middle options" (p. 20), they warn that "the existence/non-existence of middle options may result in significant differences for other selected response options" (p. 21) as the middle response option inherently contains a variety of response meanings (Yamada, 2010). Yamada (2010) has also found that if Japanese respondents are familiar with the material implicated by the question, it is preferable not to include a middle response option.

In general, there is very little research which has explored the effect of scale size on CEFR can-do statement self-assessment and even less specific to a Japanese context. In terms of the CEFR-J's development, and particularly throughout the developers' validation phases, several response scales were employed when Japanese university students were administered can-do statements (see Negishi et al., 2013). The first major survey had participants rating can-do statements for difficulty on a can-do/can't do dichotomous scale (Negishi, 2012), whereas a subsequent validation study of participant self-assessment employed a four-point Likert scale (Tono & Negishi, 2012). It should be noted that neither of these rating scales included a neutral or middle response option, probably because the illustrative descriptors were deliberately designed to be familiar to Japanese English learners. Although the developers' decision to use varying response scales is probably advantageous, scales of other sizes from those employed throughout the development process may have produced differing results, particularly in terms of the perceived difficulty of individual statements. Furthermore, for current or future practical users, whether the tendency to select a neutral response occurs when responding to CEFR-J's can-do statements remains unknown. To explore the effect of differing response scales on the outcome of a can-do statement self-assessment survey, the study reported here was designed. Specifically, Japanese university students self-assessed on the CEFR-J's A level can-do statements using both four-point and five-point scales to determine firstly, if there are any differences in response structures between the two rating scales, and secondly, if the tendency to select a neutral response exists or if the inclusion of such an option appears to affect participants' responding behaviour. Such information can be used to aid institutions, teachers or individual language learners in determining which particular response scale they might employ in their future practice, and what the effects of selecting such a scale may be.

Method

Participants

A total of 57 first year students from a private university in Western Japan participated voluntarily in this study. Students were in one of two English classes and had completed one full semester of twice-weekly ninety-minute English classes. Participants were unfamiliar with performing self-assessment using can-do statements and the CEFR-J.

Instrument

Participants indicated the extent of their agreement to 50 randomly ordered Japanese can-do statements from the CEFR-J's five A sub-levels (available for free download at http://www.cefr-j.org/english/index-e.html) on a four-point or a five-point Likert scale. Whether they responded to a four-point scale (28 participants) or a five-point scale (29 participants) was randomly determined. For the four-point scale, the categories were *Strongly Disagree*, *Disagree*, *Agree* and *Strongly Agree*. For the five-point scale, a mid-point of *Neither Agree nor Disagree* was also included. The survey was administered using SurveyMonkey®, an online data collection application ("SurveyMonkey," 2012).

Analysis

Two analyses were performed. The first was to determine whether the rating scale had any impact on mean responses, and for which can-do statements differences were apparent. The second was to explore whether Japanese self-assessors exhibited a tendency to select a neutral response or if the inclusion of a neutral response had any impact on other response selections. To test these hypotheses, individual ratings for each can-do statement were measured for each scale. Due to the unequal number of categories in each scale, the scores were made equivalent before being compared. A simple proportional transformation was used to equate the scales such that each four-point score was multiplied by 5/4 to scale it up to be equivalent to scores from the five-point scale (Colman, Norris, & Preston, 1997). It should be noted that the assumption is not that by performing the transformation, the scales become equivalent to each other. The equating process did not employ item response theory (IRT) and is thus incapable of adjusting test scores for individual test takers. As noted by Colman et al. (1997), "how people respond to rating scales with unequal numbers of response categories is a quintessentially psychological rather than a mathematical question" (p. 357). The proportional transformation was performed to determine whether the existence of a mid-point on a scale affected the behaviour of test takers, which is indeed a psychological question. Simple mathematical equating processes such as a proportional transformation, do however, allow for basic comparisons of mean scores (Livingson, 2004).

To determine if there were any main effects from response scales, the scores from all CEFR-J can-do statements were compared across the entire A-level (that is, across all five skills and all five levels) via an Analysis of Variance (ANOVA). Further ANOVAs were performed within each of the five language skills across all five CEFR-J levels. If any differences were found, LSD post-hoc tests were used to determine which can-do statements resulted in significantly different mean responses across rating scales. Statements for which significant differences exist were subsequently highlighted and analysed. Interactions between rating scales and language skills were also tested for, to observe whether the rating scale had a stronger impact on responses to can-do statements from certain language skills or levels. Finally, to determine if using a four-point scale reduces the tendency to choose a neutral response (seeing as there is no naturally neutral or middle point to select), the absolute difference between the mathematical mid-point of each scale and the response for each statement from each participant was calculated, equated by multiplying the four-point scale by 5/4 (to eliminate the difference in scale size) and then compared across scales with an ANOVA. PASW Statistics (version 18) was employed for the analyses.

Results

Both a rating scale analysis and a neutral response selection analysis were performed. The former looked for significant differences between mean responses on a four-point scale and a five-point scale to CEFR-J A-level can-do statements; and the latter looked at whether the existence of an inherent mid-point resulted in greater selection rates of that response option. In describing the results, it should be noted that a lower rating represents a higher level of difficulty whereas a higher rating reflects a lower level of difficulty.

Rating Scale Analysis

An ANOVA revealed that there was no significant main effect for rating scales across all skills and levels: overall, there were no significant differences in mean responses between the four-point scale and five-point scale (*df*= 1, *F* =.054, *p*=.817). However, when can-do statements were analysed individually within each skill and level, it was revealed that a total of eight statements exhibited significant differences between mean responses on each scale. These eight can-do statements (or 16% of the total number of statements) are shown in Table 1. Of these, two are from spoken interaction, five are from spoken production and one is from writing. Participants rated all of these eight statements as significantly higher on the five-point scale than on the four-point scale following the control to adjust for scale size.

Despite the lack of a significant main effect for rating scales, a significant interaction for rating scales and skills was found. There were significant differences between responses for reading and spoken production depending on which rating scale was employed. Specifically, reading was rated overall significantly lower (*df* =1, *F* =15.461, *p* = .000) whereas spoken production was rated significantly higher (*df* =1, *F* =6.804, *p* =.009) on the five-point scale, when compared to the four-point scale. No differences for the skills of spoken interaction, writing and listening were found.

A significant interaction for rating scales and levels was also found, such that when using the five-point scale, statements from A1.1 were being rated significantly lower (*df* =1, *F* = 15.593, *p* = .000), whereas statements from A2.2 were being rated significantly higher (*df* = 1, *F* = 5.306, *p* = .022) when compared to results from the four-point scale. No significant differences were found across rating scales for levels A1.2, A1.3 or A2.1. A three-way interaction between rating scales, skills and levels was insignificant.

Neutral Response Selection Analysis

When the overall responses were analysed to determine participants' tendency to select a neutral response, it was found that there were no differences between mean responses. For the five-point scale, a raw mean of 3.27 was selected (where the midpoint is the third category option, *Neither Agree nor Disagree*, or 3). On the four-point scale, where the theoretical midpoint is halfway between the second and third categories (or between the *Disagree* and *Agree* options, at 2.5), the raw mean response was 2.62. When adjusted for scale size using a simple transformation, these two means are equal to each other (hence the lack of main effect for rating scales), meaning that inclusion of a mid-point had no effect on the mean.

Table 1. Can-do statements exhibiting significantly differing responses between the 4-point and 5-point scales

Level	Skill	Descriptor
A2.2	Spoken Interaction	I can exchange opinions and feelings, express agreement and disagreement, and compare things and people using simple English.
A2.2	Spoken Interaction	I can interact in predictable everyday situations (e.g., a post office, a station, a shop), using a wide range of words and expressions.
A1.2	Spoken Production	I can give simple descriptions e.g. of everyday object, using simple words and basic phrases in a restricted range of sentence structures, provided I can prepare my speech in advance.
A1.3	Spoken Production	I can describe simple facts related to everyday life with a series of sentences, using simple words and basic phrases in a restricted range of sentence structures, provided I can prepare my speech in advance.
A1.3	Spoken Production	I can express simple opinions about a limited range of familiar topics in a series of sentences, using simple words and basic phrases in a restricted range of sentence structures, provided I can prepare my speech in advance.
A2.1	Spoken Production	I can give a brief talk about familiar topics (e.g. my school and my neighbourhood) supported by visual aids such as photos, pictures, and maps, using a series of simple phrases and sentences.
A2.2	Spoken Production	I can give an opinion, or explain a plan of action concisely giving some reasons, using a series of simple words and phrases and sentences.
A2.2	Writing	I can write my impressions and opinions briefly about what I have listened to and read (e.g. explanations about lifestyles and culture, stories), using basic everyday vocabulary and expressions.

To determine the extent to which a neutral response was selected for individual statements across the two scales, the absolute value of the difference between the mid-point of the scale and the actual response were compared for the four and five-point scales, following adjustments to equate the scales. It was found that a total of sixteen statements (or 32% of the total) exhibited significant differences in distance between the response and the mid-point of the scale. Of those sixteen, for fourteen of them (or 28% of the total), the distance between the scale's mid-point and the actual response was significantly greater on the five-point scale than on the four-point scale. This suggests that participants sometimes selected categories that were further away from the mid-point when a mid-point was included compared to when it was not, even after controls to adjust for scale-size.

Discussion

The results suggest that overall, rating scales did not impact greatly on Japanese university students' responses to the CEFR-J's A-level can-do statements. Only 16% of statements exhibited significant differences across rating scales (Table 1) and responses to less than one third of statements differed in terms of the distance between the response and the mid-point of the scale, illustrating that while the inclusion of a mid-point did appear to influence behaviour, its influence was neither universal nor consistent. Regarding the eight statements for which significant differences across rating scales existed, all consisted of productive language tasks (two from spoken interaction, five from spoken production and one from writing). Japanese learners of English have been shown to be more proficient in receptive language skills (reading and listening) than productive skills (speaking and writing) while also self-assessing less consistently on skills in which they are less proficient (Butler, 2004; Parry, 2000). It is perhaps for this reason that significant differences existed for a small number of productive skill can-do statements and that there were no differences between receptive skill descriptors. This possibility is also supported by the two significant interactions showing that both reading and A1.1 level statements were rated as more difficult, and spoken production and A2.2 can-do statements were rated easier, on the scale with a mid-point. This echoes Yamada's (2010) findings regarding familiarity and response scale whereby exclusion of a mid-point is advantageous on familiar items. On the other hand, both spoken production and A2.2 level tasks, were rated as less difficult on the scale with a mid-point. These findings appear to suggest that inclusion of a mid-point on a rating scale may be disadvantageous for both easy and familiar tasks as well as for difficult and unfamiliar tasks. Inclusion of a mid-point resulted in comparatively higher levels of difficulty being selected for easy and familiar tasks and lower levels of difficulty on difficult or unfamiliar tasks when compared to the responses on a scale without a mid-point. However, there were no significant differences for three out of the five language skills (listening, spoken interaction and writing), and for three out of the five CEFR-J levels (A1.2, A1.3, A2.1). Nonetheless, these findings reflect those of Murata and Onodera (2011), who found that the inclusion of a middle response option did not affect "the balance between the extreme opposite response options" (p. 21) between differing scales: the more difficult ratings on the easy or more familiar tasks were balanced out by the less difficult ratings on the more difficult or less familiar tasks.

Regarding the assumption that Japanese students will tend to choose a neutral response if given the option, the current findings are not entirely in support given that for nearly two-thirds of all responses, no tendency or preferences were observed. For the remaining third of responses, it is evident that on the scale with the mid-point, participants tended to choose responses slightly above the midpoint, whereas on the scale with no neutral option, they tended to choose the responses slightly below the mathematical mid-point. Perhaps the mere existence of a neutral option gave respondents more confidence in indicating higher mastery of the task. Alternatively, the response patterns could be related to the number of categories ahead of the selected response: for both scales, there were essentially two categories (*Strongly Agree* and *Agree*) before the mean selected response. In accordance with the previous findings about modesty (Matsuno, 2009), the possibility here is that Japanese survey-takers may not necessarily select a neutral response, but on average will select a point on the scale which does not correspond to mastery of a task, and reflects a given position on the scale which is related to the number of other options.

Overall, Japanese students' self-assessment ratings appear to be subject to a number of different factors and the findings of this study have raised a number of questions: will modesty apply more strongly for language skills that learners feel they are better at, resulting in them self-assessing more stringently? Alternatively, does added confidence in their skill also impact the surety of their response, causing them to self-assess more accurately for tasks in skills that they are more comfortable performing? Given the present results combined with previous findings, these are both possibilities. Addressing the limitations of the current study would certainly lead towards determining more specifically the factors that influence Japanese self-assessment and therefore, which type or size of rating scale is more advantageous to employ. For instance, although there were no differences in the distance between responses and mid-point when the responses from each scale were equated and compared, this was probably because absolute differences (and not whether it was over or under the mid-point) were analysed. Perhaps most importantly, however, is that IRT should be employed to determine the effects of scale size on response structures. In addition, employing a wider range of rating scales would reveal more about the impact of scale size for Japanese self-assessors, such as in Lee, Jones, Mineyama, and Zhang (2002) who found that seven response options (which contains a mid-point) had the greatest construct validity for Japanese survey-takers. Analysing the results of a greater number of participants would naturally also provide more conclusive evidence. Testing the differences between difficulty distributions across skills and within a single learner is also necessary to confirm the hypothesis that differing skills are subject to different factors when self-assessing. Lastly, employing controls for ability is important, especially since English language competence affects response styles (Harzing, 2006).

Conclusions

The findings reported here do not support the notion that Japanese survey-takers will select a neutral response if one is available. Variation in response structures depended on rating scale, skill (whether it is a productive or receptive skill) and CEFR-J difficulty level (whether it is an easy or difficult statement) but the level of variation was not consistent across all skills and levels. This suggests that familiarity with the task and the overall number of categories in the scale are more likely to impact response structures than the inclusion of a neutral response option. This is in line with the suggestion of Murata and Onodera (2011) that response structures may be affected by a variety of factors which may be survey-specific (question content, order effects, number and form of response options), item-specific (whether it was a productive or receptive language task, or familiarity with the task), socio-cultural (modesty, social desirability, resistance to response) or personal (response styles, language proficiency). It also seems likely that such variation may be eliminated following self-assessment training (Little, 2006; Rolheiser & Ross, 2013), although further exploration of the relationship between self-perception of ability, self-assessment and ability is required (Runnels, 2013).

The practical implication of this study for use of the CEFR-J is that it cannot be assumed that a learner's self-assessment will remain consistent across the skill or level being self-assessed. For educators who opt to use the CEFR-J's can-do statement scales as they are, or for students or teachers who create localized self-assessment instruments; experimentation with the rating scale is required. The present results do not determine the superiority of rating scales with or without neutral categories, but they do suggest that rating scales will, to some extent, affect response patterns and that eliminating the midpoint of a scale, as suggested by previous research, may be premature.

About the author

Judith Runnels is a graduate research student at the Centre for Research in English Language Learning and Assessment (CRELLA) of the University of Bedfordshire. Her research interests lie in assessment and evaluation and usage of the Common European Framework of Reference (CEFR).

References

Bärenfänger, O., & Tschirner, E. (2008). Language educational policy and language learning quality management: The Common European Framework of Reference. *Foreign Language Annals, 41*, 81-101.

Butler, Y. G. (2004). What level of English proficiency do elementary school teachers need to attain to teach EFL? Case studies from Korea, Taiwan, and Japan. *TESOL Quarterly, 38*(2), 245–278.

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Report, 80*, 355-362.

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

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

Faez, F., Majhanovich, S., Taylor, S., Smith, M., & Crowley, K. (2011). The power of "Can Do" statements: Teachers' perceptions of CEFR-informed instruction in French as a second language classrooms in Ontario. *The Canadian Journal of Applied Linguistics, 14*(2), 1-19.

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. Psaltou-Joycey & M. Mattheoudakis (Eds.), Advances in Research on Language Acquisition and Teaching: Selected Papers (pp. 15-26). Athens: Greek Applied Linguistics Association. Retrieved from http://www.enl.auth.gr/gala/14th/Papers/Invited%20Speakers/Fulcher.pdf

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

Harzing, A. (2006). Response styles in cross-national survey research: A 26-country study. *Journal of Crosscultural Management, 6*(2), 243-266.

Horiguchi, S., Harada, Y., Imoto, Y., & Atobe, S. (2010). The implementation of a Japanese version of the " European Language Portfolio – Junior version" in Keio: Implications from the perspective of organizational and educational anthropology. In M. A. Schmidt, N. Naganuma, F. O'Dwyer, A. Imig, & K. Sakai (Eds.), *Can do statements in language education in Japan and beyond: Applications of the CEFR* (pp. 138-154). Tokyo: Asahi Press.

Hsiao, Y. P., & Broeder, P. (2013). Applying the writing scales of the Common European Framework of Reference for Languages to the new HSK test of proficiency in Chinese: Realities, problems and some suggestions for Chinese language teachers and learners. *Language Learning in Higher Education, 2*(1), 59-74.

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

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

Kizman, A., & Nitta, K. (2010). Improving the English learning environment for Kinki University students. *English Ministry of Education Journal, 10*(6), 101-117.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a likert scale. *Research in Nursing & Health, 25*(4), 295-306.

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.

Livingson, S. (2004). *Equating test scores (without IRT)*. Princeton, N.J.: Educational Testing Service.

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

Murata, H., & Onodera, N. (2011). *Characteristics of response tendency in mail surveys: Comparing mail and face-to-face surveys. NHK Monthly Report on Broadcast Research*. Retrieved from http://www.nhk.or.jp/bunken/english/reports/pdf/report_111201-2.pdf

Nagai, N. (2010). Designing English curricula and courses in Japanese higher education: Using the CEFR as a guiding tool. In M. A. Schmidt, N. Naganuma, F. O'Dwyer, A. Imig, & K. Sakai (Eds.), *Can Do statements in language education in Japan and beyond: Applications of the CEFR* (pp. 86-104). Tokyo: Asahi Press.

Nakano, M., Tsutsui, E., & Kondo, Y. (2010). *Bridging a gap between L2 research and classroom practice (1): English as a Lingua Franca (ELF) in Asia and some assessment based on Common European Framework of Reference for Languages (CEFR)*. Paper presented at the INTERSPEECH 2010 Satellite Workshop on Second Language Studies, September 22-24, 2010, Waseda University, Tokyo, Japan. Retrieved from http://www.gavo.t.u-tokyo.ac.jp/L2WS2010/papers/L2WS2010_P1-02.pdf

Negishi, M. (2011). The development process of the CEFR-J [in Japanese]. *ARCLE Review, 5*(3), 37-52.

Negishi, M. (2012). The Development of the CEFR-J: Where we are, where we are going. In N. Tomimori, M. Furihata, K. Haida, N. Kurosawa, & M. Negishi (Eds.), New perspectives for foreign language teaching in higher education: Exploring the possibilities of application of CEFR (pp. 105-116). Tokyo: WOLSEC, Tokyo University of Foreign Studies. Retrieved from http://www.tufs.ac.jp/common/fs/ilr/EU_kaken/_userdata/negishi2.pdf.

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. (2007). The CEFR Common Reference Levels: Validated reference points and local strategies. In F. Goullier (Ed.), *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: Challenges and responsibilities. Report of the Council of Europe Intergovernmental Language Policy Forum (Strasbourg, France, 6-8 February 2007)*. Strasbourg, France: Council of Europe, Language Policy Division.

O'Dwyer, F. (2013). *The Use of the CEFR and can-do statement*. Paper presented at the Hiroshima JALT Chapter Meeting Workshop, 19th September 2013, Hiroshima, Japan. Retrieved from http://www.academia.edu/4522808/Handout_for_CEFR_Can_do_statements_Workshop_Sep_29th_2013

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

Parmenter, L., & Byram, M. (2010). An overview of the international influences of the CEFR. In M. S. Schmidt, N. Naganuma, F. O'Dwyer, A. Imig, & K. Sakai (Eds.), *Can do statements in language education in Japan and beyond* (pp. 9-17). Tokyo Asahi Press.

Parry, M. (2000). How 'communicative' are introductory undergraduate level Japanese language textbooks. *Japanese Studies, 20*(1), 89-102.

Pashupati, K., Courtright, M., & Pettit, A. (2013). *Examining respondent scale usage across 10 countries*. Paper presented at the 2013 CASRO Online Research Conference, March 7-8, 2013, San Francisco.

Porto, M., & Barboni, S. (2012). Policy perspectives from Argentina. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalization of language education policy*. Bristol: Multilingual Matters.

Rolheiser, C., & Ross, J. (2013). Student self-evaluation: What research says and what practice shows. Retrieved February 10th, 2013, from http://www.cdl.org/resource-library/articles/self_eval.php

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

Ryan, S. (2009). Self and identity in L2 motivation in Japan: The ideal L2 self and Japanese learners of English. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 120-143). Bristol: Multilingual Matters.

Semmelroth, A. (2013). Aligning a language curriculum to the CEFR-J. *Framework and Language Portfolio SIG Newsletter, 10*, 6-19.

SurveyMonkey. (2012). from Surveymonkey.com, LLC. http://www.surveymonkey.com

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

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.

TUFS Tonolab. (2012). *CEFR based framework for ELT in Japan*. Retrieved from http://www.tufs.ac.jp/ts/personal/tonolab/cefrj/english/download.html

Wang, H., Kuo, B., Tsai, Y., & Liao, C. (2012). A CEFR-based computerized adaptive testing system for Chinese proficiency. *The Turkish Online Journal of Educational Technology, 11*(4). Retrieved from http://www.tojet.net/articles/v11i4/1141.pdf

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

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

Wu, J. (2012). Policy perspectives from Taiwan. In M. Byram & L. Parameter (Eds.), *The Common European Framework of Reference: The globalization of language education policy* (pp. 213-223). Bristol: Multilingual Matters.

Yamada, K. (2010). *The skill of asking: Guidelines for creating questionnaires for research*. Tokyo: Nikkei Publishing Inc.

Yoneka, J. (2011). From CEFR to CAFR: Place for a Common Asian Framework of Reference for languages in the East Asian business world? *Asian Englishes, 14*(2), 86-91.