# An Exploratory Reliability and Content Analysis of the CEFR-Japan's A-Level Can-Do Statements

**Judith Runnels**
*Hiroshima Bunkyo Women's University*

Both the Common European Framework of Reference (CEFR) and the CEFR-Japan (CEFR-J), an alternate version designed for Japanese learners of English, provide measurements of language proficiency via assessment or self-assessment on scales of descriptors of communicative competences (known as can-do statements). Although extensive empirical evidence supports these claims for the CEFR, the same cannot yet be said of the CEFR-J. Mokken scaling was thus used to measure the reliability of can-do statement scales from the five skills of the CEFR-J's five A sublevels of A1.1, A1.2, A1.3, A2.1, and A2.2. Statements that negatively affected the reliability of the scale were analysed. Lower reliability was attributed to characteristics specific to participants (homogeneity of the population, familiarity with the task, and if the material was recently studied), and content of the statement itself (whether it implied more than one language skill or none at all, whether it contained a contradiction, or was confusing or unfamiliar). Modifications to increase the reliability of can-do statement scales and limitations of using illustrative descriptor-based systems as measurement instruments are discussed.

ヨーロッパ共通言語参照枠（CEFR）とその日本版CEFR-Jはともにコミュニケーション能力を示す指標（can-do statements）であり、評価あるいは自己評価による言語運用能力の測定を目的としている。CEFRにはその主張を裏付ける根拠が豊富にある一方で、CEFR-Jには未だ十分な裏付けがあるとは言い難い。そこで本研究は、CEFR-Jの5つのＡ sublevelについて5技能に関わるcan-do statementの信頼性をMokken スケールを用いて測定した。信頼性に否定的な影響を与えた指標をさらに分析したところ、信頼性の低さは学習者特有の特徴（母集団の均一性、課題に対する慣れ、最近学習された項目か否か）と指標そのもの（2つ以上の言語技能に関係している、言語技能に関係していない、矛盾がある、あいまいでわかりにくい）に起因するものであっ

た。Can-do statementsの信頼性を高めるための修正に関する提案と、ディスクリプタを用いた指標を使用することの限界についての考察を行った。

Theoretical work, case studies, and other evidence suggest that the Common European Framework of Reference (CEFR) provides an effective descriptive scheme for analysing the needs, goals, materials, and achievements of language learners (Alanen, Huhta, & Tarnanen, 2010; Council of Europe, 2001). It employs illustrative descriptors, known as can-do statements, of communicative competences for five skills (listening, reading, spoken interaction, spoken production, and writing). All can-do statements are divided into six proficiency levels of increasing difficulty (A1, A2, B1, B2, C1, C2). To provide an example, can-do statements 1 and 2 are from reading levels B1 and A1 respectively:

1. *I can identify the main conclusions in clearly written argumentative texts*.
2. *I can understand the general idea of simple informational texts and short simple descriptions, especially if they contain pictures which help to explain the text*. (Council of Europe, 2001)

The CEFR's descriptors were developed through qualitative and quantitative methods to ensure progressions in difficulty as a learner advances through the levels (North, 2000; 2002). This difficulty hierarchy has been continually validated in a European context since the CEFR's publication (Figueras, 2012). Although the CEFR is argued to be an "international standard for language teaching and learning" (North, Ortega, & Sheehan, 2010, p. 6), it is also frequently criticized for its theoretical underpinnings, particularly regarding how it should be used to measure proficiency. Because the hierarchy of difficulty represented by the increasing levels is largely based on difficulty judgments from language educators, Fulcher (2004; 2010) argued that it can neither be used to gauge proficiency nor provide any standardized measure of language ability. Other opponents of the CEFR have noted that it cannot and should not act as a language test for measuring ability (Weir, 2005), as ties to SLA theory have yet to be established (Hulstijn, 2007), and, as well, the progression of difficulty inherent in the levels is unsupported by empirical studies of performance samples from language learners (Westhoff, 2007).

Conversely, supporters of the CEFR praise it for how it can be used by autonomous learners to provide an estimation of proficiency or direction for an individual's language study (Glover, 2011). Typically, such a level estimation is achieved with a self-assessment whereby learners read a set of can-do statements then decide if they are capable of performing the communicative actions entailed by each statement (Glover, 2011; Little, 2006). Level estimations are thus based on the learners' perceptions of their own achievement on the can-do statements. Future self-assessments can be compared to previous ones as a measure of progress.

Due to its success in Europe (North, et al., 2010) and other regions of the world (Figueras, 2012; Wang, Kuo, Tsai, & Liao, 2012), the CEFR has been modified into alternate versions tailored to meet local demands. One such example is the CEFR-Japan (CEFR-J), introduced to address the lack of consistently used measures for progress or proficiency among Japanese institutions.

## Developing the CEFR-J

When Negishi (2012) found that over 80% of Japanese English learners fall within A1 and A2 levels, he concluded that the CEFR's can-do statements were not providing users with adequate criteria for distinguishing between the population's language abilities. He highlighted the need for a system tailored to the needs of Japanese English language learners and development of the CEFR-J thus began (Negishi, Takada, & Tono, 2013; Tono & Negishi, 2012). As part of the first stage of development, can-do statements from DIALANG (Council of Europe, 2001, p. 231-234; Huhta, Luoma, Oscarson, Sajavaara, Takala, & Teasdale, 2002) were administered to 360 Japanese university students to ascertain that the rank ordering of difficulty by Japanese students matched that of the CEFR (Negishi et al., 2013). The participants generally ordered the can-do statements accordingly and it was concluded that overall, the CEFR would be suitable for use by Japanese English learners. Nonetheless, there were some outlying can-do statements that were being rated by the Japanese population as more difficult than predicted. Negishi's (2011) analysis of an outlying A1 reading descriptor is as follows:

> *I can understand short, simple messages, e.g., on postcards* turned out to be more difficult than the A2.1 descriptor *I can understand short, simple texts containing the most common words, including some shared international words*. This might be because Japanese postcards tend to contain much more information than their European counterparts, and therefore the Japanese EFL learners considered it to be more difficult than it was originally assumed in the CEFR. (p. 108)

Negishi (2011) concluded that tasks "were judged to be more difficult than the levels they were originally assigned to [if learners had only had limited experience with them], whereas the tasks they had experienced were judged to be easier" (p. 108). Any can-do statement that was not scaling according to the CEFR was thus adjusted with real-life examples specific to a Japanese context and then retested. Following modifications, the initially outlying can-do statements ordered consistently with the CEFR's predictions, thus demonstrating that the contextualization or localization process had been successful (Negishi et al., 2013).

In addition to the contextualization of descriptors, the CEFR's A and B levels were modified in order to better distinguish between learners (Negishi, 2011; Tono & Negishi, 2012). The CEFR's four original levels (A1, A2, B1, B2) were subdivided into nine categories (A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2) and a Pre-A1 level was also created, resulting in 12 CEFR-J levels in total. Since its publication in March 2012 (TUFS Tonolab, 2012), the CEFR-J has been promoted as a way forward for any language education program in Japan and numerous projects for developing textbooks and learner and teacher autonomy support tools are under way (Imig, 2013).

## Reliability Issues

Despite interest in the implementation of the system, only a limited amount of research specific to the CEFR-J has been undertaken. For the CEFR, extensive work has demonstrated the reliability of a scale of increasing difficulty, which in turn supports any arguments regarding standardized assessment of language proficiency (Little, 2006; North, 2007; North & Schneider, 1998). For the CEFR-J however, very few studies beyond the CEFR-J's development process (Negishi, 2011; Negishi et al., 2011; Tono & Negishi, 2012) have been published. Runnels (2013a) for instance, measured the rank ordering of can-do statement difficulty by Japanese university-level English language learners for the CEFR-J's levels A1.1 through A2.2. Rasch analysis (Andrich, 1978) and analyses of variances (ANOVAs) indicated that for several adjacent levels, there were no statistically significant differences in mean difficulty ratings. This was also found to be the case when differences between levels within each individual skill were tested (Runnels, 2013b). Although Tono and Negishi (2012) had concluded that the two original CEFR A levels were not adequately distinguishing among the span of learners' abilities, Runnels (2013a, 2013b) suggested that the five sublevels the A level was divided into were perhaps too many and that the support of the language learning process that the CEFR-J is designed to provide may be jeopardized if users cannot consistently distinguish between levels of proficiency based on the CEFR-J descriptors.

A major weakness of both of Runnels' (2013a, 2013b) studies, however, is that they were focused solely on the responses to the difficulty of can-do statements from a limited sample of users and did not provide any measure of reliability or variance to account for individual differences across participants. In general, when difficulty ratings alone are analysed, the extent to which the difficulty hierarchy may be different for every learner is not accounted for. Although an issue certainly exists if the CEFR-J's users are not able to confidently estimate language level due to negligible differences in difficulty between adjacent sublevels, there is also an issue if CEFR-J can-do statements and their scales are behaving very differently for each individual that responds to them. For instance, learners who are using an A2.2 can-do statement such as 3 to self-assess listening may conclude that they are able to perform any task entailed by 3. However, the next learner to self-assess using 3 may not come to the same conclusion, deciding instead that he or she can only perform tasks from 4 or 5, both lower order statements than A2.1:

3. *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*.
4. *I can understand short, simple announcements (e.g., on public transport or in stations or airports), provided they are delivered slowly and clearly.*
5. *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.*

Assuming that controls for rater severity and ability are taken into account, placing the first learner at A2.2 for listening and the second at A2.1 level is supported by empirical demonstrations of common understanding of the difficulty of statements across populations of users. For the CEFR-J, though, these conclusions cannot be drawn with such confidence because no prior researcher has examined whether A2.2 can-do statements such as 3 are indeed rated as more difficult than A2.1 descriptors such as 4 and 5 by the majority of users. Empirical studies demonstrating a consistent and reliable difficulty hierarchy across the levels of the CEFR-J are lacking.

The current study was thus designed to provide preliminary evidence on the reliability of can-do statements within the CEFR-J's difficulty hierarchy and to determine the extent to which participants are behaving consistently in their responses regarding the difficulty of can-do statements within each skill's scale. Any A-level can-do statement shown to be negatively affecting the reliability of a skill's scale (in that response patterns are found to be less consistent) is analysed, and recommendations for modification in order to potentially increase reliability are discussed.

## Method

### *Participants*

Participants consisted of 590 first- and 2nd-year students from a small private women's university in western Japan. Each participant was in one of five majors of study, one of which was English. In order to determine whether the can-do statements were interpreted consistently across a variety of users with a range of language abilities, both 1st- and 2nd-year students from all of these disciplines were included in the analysis. The can-do statement survey described below was administered at the end of the first semester of the academic year, meaning that all non-English majors (536 participants or 90.8% of the total) had completed at least 4 months or 12 months of twice-weekly 90-minute university-level English classes, depending on whether students were in their 1st or 2nd year of study. The English majors had completed one or three semesters of full-time English study depending on whether they were in their 1st or 2nd year.

All participants were unfamiliar with the CEFR-J and had no previous experience using can-do statements. They had also received no training on conducting self-assessment. Participation was voluntary and had no bearing on course grades.

### *Instrument*

The can-do statement survey was administered on www.surveymonkey.com (SurveyMonkey, 2012) during participants' class time and in their regular classrooms. All statements are available online in both English and Japanese from the Tokyo University of Foreign Studies (TUFS Tonolab, 2012). For each of the five skills (listening, reading, spoken production, spoken interaction, and writing) there are two can-do statements for each level, for a total of 50 statements. Participants responded on a 5-category Likert scale from *strongly disagree* to *strongly agree* to all randomly ordered Japanese can-do statements from the CEFR-J's five A sublevels (A1.1, A1.2, A1.3, A2.1 and A2.2), which were selected because the institution's curriculum is targeted at these levels.

## Analysis

Multivariate Statistics Inc.'s EQSIRT Version 1.0 (Bentler & Wu, 2012) was used to perform a Mokken Scale analysis to determine the reliability of the can-do statements' scales for each skill.

In testing, Guttman patterning is an ideal hypothetical pattern of item difficulties (Guttman, 1950). If the test forms a theoretically perfect Guttman pattern, all test-takers will reach a point in the question lineup (wherein all questions are lined up in order of increasing difficulty) such that all of the questions have been answered correctly up to that point, but all of the questions afterwards are too difficult and are therefore answered incorrectly. The point at which the change from correct to incorrect occurs depends on the test-takers and is often seen to represent their ability on that test. Mokken scaling is a statistical technique that assumes the order of difficulty of items is not the same across a population (van Schuur, 2003) and it provides a measure of reliability by identifying items for which Guttman patterning is occurring at higher rates (Molenaar, 1997; Sijtsma & Molenaar, 2002).

The CEFR-J's increasingly difficult levels or hierarchy theoretically forms a Guttman scale: A1.1 should be easier than A1.2 which is easier than A1.3 and so forth, such that learners will eventually reach a point beyond which the tasks are too difficult for them to perform, thus representing their CEFR-J level of proficiency. Accordingly, in this theoretically perfect system, learners should find A2.1 listening statements such as 4 and 5 easier than A2.2 statement 3. Realistically, this may not always be the case as some learners will find A2.1 statements more difficult. The response patterns from these learners would thus contradict the intended Guttman patterning of the system. For example, if a particular learner finds an A2.1 item to be extremely difficult and an A2.2 item very easy, while peers of the same ability find the A2.1 item easier, the distribution of difficulty ratings for the A2.1 can-do statement would then skew, with the mean difficulty rating increasing due to the responses from only a few learners even though the majority of respondents of the same ability were behaving similarly to one another.

Mokken scaling detects for these types of response patterns by creating a scale that reflects the difficulties of each statement according to the abilities of respondents, but also the extent to which a greater number of more able respondents found the given statement more difficult (van Schuur, 2003). Its resulting statistic, known as the coefficient of homogeneity (*H* or *H*-value), reflects response structures for each item in terms of item thresholds (Andrich, 1978; Embretson & Reise, 2000) and provides a measure of reliability for each can-do statement, reflecting the extent to which a Guttman pattern is evident for all responses. Coefficients of homogeneity fall between 0 and 1.0, where a higher *H*-value is associated with an item that is scaling more Guttman-like (Mokken, 1971). Unacceptable *H*-values fall below .3, and anything over .6 is considered strong in terms of reliability (van Schuur, 2003).

For the current analysis, an *H*-value provides an alternate perspective to Cronbach's alpha as a measure of reliability because "the order of 'difficulty' of the items has an important theoretical interpretation that is not taken into consideration in [traditional] reliability analyses" (van Schuur, 2003, p. 141). Although classical reliability analyses assume that all items exhibit the same frequency distributions, when items are expected to form a Guttman scale such as in the CEFR-J difficulty hierarchy, the assumption is the opposite: that items exhibit differing frequency distributions (Carroll, 1945; Ferguson, 1941; van Schuur, 2003). Therefore, "if items in fact form a Guttman scale, or are expected to do so, it makes sense to analyse them with a model that takes Guttman's model assumption of cumulativity into account" (van Schuur, 2003, p. 141).

## Results and Discussion

A Mokken Scale analysis, performed to examine the reliability of all A-level can-do statements, revealed that the CEFR-J's A1- and A2-level can-do statements are forming a strongly reliable scale (*H* = .624) according to commonly accepted criteria for *H* (van Schuur, 2003). Cronbach's alpha was found to be .944 across all statements, also indicating that overall, the scales were found to be strongly reliable.

The results of the Mokken Scale analysis for each statement are displayed in Tables 1-5 according to language skill. The *H*-value next to each statement represents the reliability of the scale as a whole. If a given statement is removed, there is either a positive (moving down in the table) or negative (moving from the bottom up) impact on the reliability of the scale. Therefore, statements that are closer to the top of the table are more strongly affecting the reliability of the scale in a negative way. Of particular concern are any statements from higher order CEFR-J levels that are appearing near the top of the table, because they are theoretically more difficult to perform and should therefore appear further down in the table as a result of being rated more difficult by a larger number of respondents. The two statements at the bottom of each scale exhibit the same coefficient of homogeneity because the *H*-value is incomputable for less than three items (i.e., three items are required to constitute a scale). For each skill, the least reliable statements will be further analysed in terms of how their content may be affecting reliability. Specifically, four descriptors from Listening (L), three from Reading (R), two from Spoken Production (SP) and Writing (W), and one statement from Spoken Interaction (SI) are discussed. Of these, four are from A1.1 and A2.1, three are from A2.2, and one is from A1.3.

**Table 1. Mokken Scales for the CEFR-J A-Level Can-Do Statements for Listening**

| Rf | Level | *H* | Listening can-do statement |
|---|---|---|---|
| (a) | A2.1 | .64 | I can understand short, simple announcements (e.g., on public transport or in stations or airports) provided they are delivered slowly and clearly. |
| (b) | A2.2 | .66 | I can understand and follow a series of instructions for sports, cooking, etc. provided they are delivered slowly and clearly. |
| (c) | A1.3 | .67 | I can understand instructions and explanations necessary for simple transactions (e.g., shopping and eating out), provided they are delivered slowly and clearly. |
|  | A1.1 | .68 | I can understand short, simple instructions such as “Stand up.” “Sit down.” “Stop.” etc., provided they are delivered face-to face, slowly and clearly. |
|  | A1.1 | .69 | 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. |
|  | A2.1 | .70 | 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. |
|  | A1.2 | .71 | I can understand short conversations about familiar topics (e.g., hobbies, sports, club activities), provided they are delivered in slow and clear speech. |
|  | A1.2 | .72 | 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. |

| Rf | Level | *H* | Listening can-do statement |
|---|---|---|---|
| (d) | A2.2 | .74 | 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. |
| | A1.3 | .74 | 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. |

*Note. H*-values represent the reliability of the scale as a whole. Statements closer to the top more strongly affect the reliability of the scale in a negative way.

The two least reliable listening items in Table 1, references (a) and (b), are both A2-level statements. The less reliable responses to (a) may be attributable to participants' lack of experience with English-language announcements as was found in Negishi (2011), whereby familiar tasks were judged as linguistically easier to complete than nonfamiliar tasks.

Regarding the comprehension of short public announcements, however, given that many in train stations in Japan announcements are made bilingually (in Japanese and English), participants may not normally rely on the English announcement to obtain information they need as the first part of the announcement is typically in Japanese, with the English following. It may therefore be difficult for participants to conceive of their performance on this task given no real prior experience. Alternatively, the inconsistent responses to this can-do statement may have been subject to a contradiction contained within: stations or airports are typically loud and busy places, and announcements in such places are not likely to be delivered slowly and clearly.

Item (b) (from A2.2) also decreases the scale's reliability. When this statement is compared with the more reliable A2.2 statement, (d), it is evident that the latter includes greater detail regarding the circumstances surrounding performance of the task despite nearly identical content. This suggests that can-do statements may scale more reliably if the criteria of the can-do statement is more specific in that contextual and performance details of the task are provided (Green, 2012). However, it is unclear as to why these statements are at the same difficulty level when (d) appears to provide considerably more support to the listener than (b).

The third least reliable L descriptor is from level A1.3, item (c). Although the task entailed by this statement is deemed an L task, perhaps its lower reliability is due to the implication that spoken interaction is also required. Although the statement does not explicitly require a response in navigating the transaction, participants may not have considered this to be solely a listening task. Respondents may also have been confused about what kind of instructions or explanations are involved when shopping or eating out. If this can-do statement refers to listening to how products are made or how food is prepared, the difficulty of language required for that level of comprehension is likely much higher than A1.3.

**Table 2. Mokken Scales for the CEFR-J A-Level Can-Do Statements for Reading**

| Rf | Level | *H* | Reading can-do statement |
|---|---|---|---|
| (e) | A1.1 | .63 | I can understand a fast-food restaurant menu that has pictures or photos, and choose the food and drink in the menu. |
| (f) | A1.1 | .64 | I can read and understand very short, simple, directions used in everyday life such as “No parking”, “No food or drink”, etc. |
| | A2.2 | .66 | I can understand short narratives and biographies written in simple words. |
| | A1.3 | .67 | I can understand short narratives with illustrations and pictures written in simple words. |
| | A2.1 | .68 | I can find the information I need, from practical, concrete, predictable texts (e.g., travel guidebooks, recipes), provided they are written in simple English. |
| | A1.2 | .69 | I can understand very short reports of recent events such as text messages from friends’ or relatives’, describing travel memories, etc. |
| | A1.2 | .7 | I can understand very short, simple, everyday texts, such as simple posters and invitation cards. |
| | A2.2 | .72 | I can understand the main points of texts dealing with everyday topics (e.g., life, hobbies, sports) and obtain the information I need. |

| Rf | Level | *H* | Reading can-do statement |
|---|---|---|---|
| | A2.1 | .74 | I can understand explanatory texts describing people, places, everyday life, and culture, etc., written in simple words. |
| | A1.3 | .74 | I can understand texts of personal interest (e.g., articles about sports, music, travel, etc.) written with simple words supported by illustrations and pictures. |

*Note. H*-values represent the reliability of the scale as a whole.

In terms of the reliability of the reading scales (Table 2), tasks entailed by the least reliable statement, (e), do not seem to involve reading. Completion of this task could simply involve looking at photographs then pointing and nodding—behavior that is independent of language ability. This may also be the case for the second-least reliable statement, (f), also from A1.1. The examples in (f) are often presented graphically or concurrently with graphics and the directions may be comprehensible without reading. The reliability of this statement may increase if other short, simple directions that are not associated with images were included, thus rendering it a strictly reading task rather than picture-identification. Nonetheless, these statements do appear at the bottom of the Mokken Scale, in accordance with the CEFR-J's difficulty hierarchy, and in this sense, the responses to these statements were as expected.

## Table 3. Mokken Scales for the CEFR-J A-Level Can-Do Statements for Spoken Interaction

| Rf | Level | *H* | Spoken Interaction can-do statement |
|---|---|---|---|
| (g) | A2.1 | .62 | I can give simple directions from place to place, using basic expressions such as "turn right" and "go straight" along with sequencers such as first, then, and next. |
| | A1.1 | .64 | I can ask and answer questions about times, dates, and places, using familiar, formulaic expressions. |
| | A1.3 | .65 | I can make, accept and decline offers, using simple words and a limited range of expressions. |

| Rf | Level | *H* | Spoken Interaction can-do statement |
|---|---|---|---|
| | A1.1 | .66 | I can ask and answer about personal topics (e.g., family, daily routines, hobby), using mostly familiar expressions and some basic sentences (although these are not necessarily accurate). |
| | A1.2 | .66 | I can respond simply in basic, everyday interactions such as talking about what I can/cannot do or describing colour, using a limited repertoire of expressions. |
| | A2.2 | .67 | I can interact in predictable everyday situations (e.g., a post office, a station, a shop), using a wide range of words and expressions. |
| | A2.1 | .67 | I can get across basic information and exchange simple opinions, using pictures or objects to help me. |
| | A2.2 | .68 | I can exchange opinions and feelings, express agreement and disagreement, and compare things and people using simple English. |
| | A1.2 | .69 | I can exchange simple opinions about very familiar topics such as likes and dislikes for sports, foods, etc., using a limited repertoire of expressions, provided that people speak clearly. |
| | A1.3 | .69 | I can ask and answer simple questions about familiar topics such as hobbies, club activities, provided people speak clearly. |

*Note. H*-values represent the reliability of the scale as a whole.

For spoken interaction, the lower reliability of the A2.1-level (g) in Table 3 could be accounted for by considering recently studied course content. Many participants rated this statement as easier than its predicted difficulty—as an A2.1-level statement, it should appear much lower in the table. One possibility is that this statement may have been considered more of a speaking skill by some participants, as giving directions could potentially entail responding to the initial request for directions rather than interacting in the traditional sense. However, half of the participants (the 2nd-year students) had recently become familiarized with completing this task whereas the other half (the 1st-year students) had little or no experience with it. For the 2nd-year student participants, three out of 30 lessons or 10% of the semester's materials were focused on giving and following directions—essentially a task derived directly from this statement. In fact, this is also the case for tasks entailed by the spoken production statements from levels A2.2 and A2.1, (h) and (i) in Table 4, as the 1st-year students had recent experience with this task, having completed four out of 30 lessons (or just over 13% of the semester) on this topic. This suggests that differences inherent in participant demographics may significantly influence scaling and that both the homogeneity of the sample and recent experiences of participants should be kept in mind when determining difficulty. These findings also reiterate the importance of performing a reliability analysis rather than a difficulty analysis alone.

**Table 4. Mokken Scales for the CEFR-J A-Level Can-Do Statements for Spoken Production**

| Rf | **Level** | *H* | **Spoken Production can-do statement** |
|---|---|---|---|
| (h) | A2.2 | .69 | I can make a short speech on topics directly related to my everyday life (e.g., myself, my school, my neighborhood) with the use of visual aids such as photos, pictures, and maps, using a series of simple words and phrases and sentences. |
| (i) | A2.1 | .70 | I can introduce myself including my hobbies and abilities, using a series of simple phrases and sentences. |
| | A1.3 | .70 | 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 | .70 | 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. |
| | A1.2 | .71 | I can give simple descriptions (e.g., of everyday objects) using simple words and basic phrases in a restricted range of sentence structures, provided I can prepare my speech in advance. |

| Rf | **Level** | ***H*** | **Spoken Production can-do statement** |
|---|---|---|---|
| | A1.1 | .71 | I can convey simple information (e.g., times, dates, places), using basic phrases and formulaic expressions. |
| | A1.2 | .72 | I can express simple opinions related to limited, familiar topics, using simple words and basic phrases in a restricted range of sentence structures, provided I can prepare my speech in advance. |
| | A1.1 | .73 | I can convey personal information (e.g., about my family and hobbies), using basic phrases and formulaic expressions. |
| | A2.2 | .76 | 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. |
| | A2.1 | .76 | I can give a brief talk about familiar topics (e.g., my school and my neighborhood) supported by visual aids such as photos, pictures, and maps, using a series of simple phrases and sentences. |

*Note. H*-values represent the reliability of the scale as a whole.

For writing, the third statement in Table 5 from A2.2, (j), negatively affects the reliability, possibly because it implicates use of a varied range of communicative competencies from W, R, and L. In this case, the reliability analysis might be highlighting the importance of unidimensionality in a can-do statement such that descriptors that implicate more than a single skill may behave less reliably.

**Table 5. Mokken Scales for the CEFR-J A-Level Can-Do Statements for Writing**

| **Rf** | **Level** | ***H*** | **Writing can-do statement** |
|---|---|---|---|
| | A1.1 | .62 | I can fill in forms with such items as name, address, and occupation. |
| | A1.1 | .64 | I can write short phrases and sentences giving basic information about myself (e.g., name, address, family) with the use of a dictionary. |

| Rf | Level | *H* | Writing can-do statement |
|---|---|---|---|
| (j) | A2.2 | .66 | 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. |
| | A2.1 | .66 | I can write texts of some length (e.g., diary entries, explanations of events) in simple English, using basic, concrete vocabulary and simple phrases and sentences, linking sentences with simple connectives like and, but, and because. |
| | A1.3 | .67 | I can write short texts about my experiences with the use of a dictionary. |
| | A2.2 | .69 | I can write a simple description about events of my immediate environment, hobby, places, and work, provided they are in the field of my personal experience and of my immediate need. |
| | A1.2 | .70 | I can write short texts about matters of personal relevance (e.g., likes and dislikes, family, and school life), using simple words and basic expressions. |
| | A1.3 | .71 | I can write a series of sentences about my hobbies and likes and dislikes, using simple words and basic expressions. |
| | A2.1 | .75 | I can write invitations, personal letters, memos, and messages, in simple English, provided they are about routine, personal matters. |
| | A1.2 | .75 | I can write message cards (e.g., birthday cards) and short memos about events of personal relevance, using simple words and basic expressions. |

*Note. H*-values represent the reliability of the scale as a whole.

## Conclusions

The reliability analysis (Tables 1-5) provided an alternate view of can-do statement scales by taking differing frequency distributions into consideration and revealing response patterns otherwise not evident if difficulty information alone is used to create a hierarchy. It was found that the can-do statements for each of the CEFR-J's A1 and A2 skills formed strongly reliable scales according to both Cronbach's alpha and Mokken Scaling's Coefficient of Homogeneity. Nonetheless, some statements negatively affected the reliability of the scale. Of particular concern are the higher level CEFR-J statements that were found close to the tops of Tables 1-5, as this reflects inconsistent difficulty ratings from a larger number of able participants.

Overall, the results indicate that the reliability of difficulty judgements on can-do statements may be affected by two main factors: the content of the can-do statement itself and specific characteristics of the population of respondents. In terms of the former, the results suggest reliability scores may be impacted by the specificity of criteria information (Green, 2012), whether the statement appeared to contain confusing or unfamiliar content, contradict itself, or imply either more than one skill or no language use whatsoever. Regarding the population of participants, reliability may be influenced by either familiarity or lack of experience with the task, whether participants had recently studied any material relevant to task performance, and the homogeneity of the population of participants.

This study provides some preliminary albeit limited findings on the reliability of the CEFR-J's A-level can-do statements and scales, suggesting that both could benefit from further empirical evidence to ensure that the system as a whole is functioning as intended. The analysis also highlights some considerations for future study. In this study, individual differences in a population of learners were shown to affect difficulty ratings and in turn, reliability scores on both the can-do statements and skill scales. Furthermore, examination of statements that were negatively affecting the reliability of the CEFR-J's skill scales suggested that content modification or adjustment in level may improve future versions of the system by increasing common understanding of the statements and their intended difficulties.

These findings have implications for future use of the CEFR-J and iterate issues associated with using can-do scales as measuring instruments for language proficiency. The importance of including checks for reliability is also emphasized, as individual learner characteristics are overlooked when difficulty ratings alone are used as the basis for creation of a scale. CEFR-J users should thus be mindful that unlike for the CEFR, which boasts significantly more supporting empirical evidence, sets of CEFR-J can-do statements may not behave identically or even similarly across and within different populations of learners. They should also be aware that estimations of levels derived from can-do instruments—via self-assessment or otherwise—may not be comparable within or across those same populations. Naturally, if task performance instead of self-assessment had been measured, differing reliability scores or response patterns might have been found. In fact, little research on the relationship between ability, self-assessment, and CEFR-aligned task performance for Japanese learners has been carried out. Further studies on this, the CEFR-J's target users' responses to can-do statements, and content analyses of the can-do statements should be performed to ensure a consistent, common interpretation of the system.

## Acknowledgments

I would like to acknowledge the contribution of Jeffrey Stewart from Kyushu Sangyo University in Fukuoka, Japan and Aaron Olaf Batty of Keio University, Tokyo. I am also indebted to Takami Takeuchi, a Hiroshima JALT member. Correspondence concerning this article should be addressed to judith.runnels@gmail.com.

**Judith Runnels** was a lecturer and head of assessment at Hiroshima Bunkyo Women's University. Her research interests are in testing and evaluation and the CEFR and its usage in learner self-assessment. She is now a postgraduate research student in the UK.

## References

Alanen, R., Huhta, A., Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 21-56). Colchester, UK: Eurosla.

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*, 665-680. http://dx.doi.org/10.1177/001316447803800308

Bentler, P. M., & Wu, E. J. C. (2012). EQSIRT 1.0 for Windows [Computer software]. Encino, CA: Multivariate Software.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1-19. http://dx.doi.org/10.1007/BF02289789

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*, 323-330. http://dx.doi.org/ 10.1007/BF02288588

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477-485. http://dx.doi.org/ 10.1093/elt/ccs037

Fulcher G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*, 253-266. http://dx.doi.org/10.1207/s15434311laq0104_4

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* (pp. 15-26). Thessaloniki, Greece: GALA.

Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness, 20*, 121-133. http://dx.doi.org/10.1080/09658416.2011.555556

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

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer et al. *(*Eds.), *Measurement and prediction. Studies in social psychology in World War II* (pp. 60-90). Princeton, NJ: Princeton University Press.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). DIALANG: A diagnostic language assessment system for learners. In J. C. Alderson (Ed.), *Common European Framework of Reference for languages: Learning, teaching, assessment. Case studies* (pp. 130-145). Strasbourg, France: Council of Europe.

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

Imig, A. (2013). Development of EAP Textbooks based on the CEFR and learner/ teacher autonomy support tools. *Framework and Language Portfolio SIG Newsletter*, *9*, 2-3.

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*, 167-190. http://dx.doi.org/ http://dx.doi.org/10.1017/S0261444806003557

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research.* New York: De Gruyter.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 367-380). New York: Springer-Verlag.

Negishi, M. (2011). CEFR-J kaihatsu no keii [The development process of the CEFR-J]. *ARCLE Review, 5*(3), 37-52.

Negishi, M. (2012). The development of the CEFR-J: Where we are, where we are going. Grant-in-Aid for Scientific Research Research Project Report (pp. 105-116). 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. (2002). Developing descriptor scales of language proficiency for the CEF common reference levels. In J. C. 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). The CEFR common reference Levels: Validated reference points and local strategies. In G. Francis (Ed.), *Report of the Intergovernmental Language Policy Forum: "The Common European Framework of Reference for languages (CEFR) and the development of language policies: Challenges and responsibilities*" (pp. 19-29). Strasbourg, France: Council of Europe.

North, B., Ortega, A., & Sheehan, S. (2010). *A core inventory for general English, British Council/EAQUALS.* Retrieved from: https://www.teachingenglish.org.uk/article/british-council-eaquals-core-inventory-general-english-0

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*, 217-262. http://dx.doi.org/10.1177/026553229801500204

Runnels, J. (2013a). Preliminary validation of A1 and A2 sub-levels of the CEFR-J. *Shiken Research Bulletin*, *17*(1), 3-10.

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

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

SurveyMonkey. (2012). Surveymonkey.com©, LLC. Palo Alto, CA: www.surveymonkey.com.

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

TUFS Tonolab. (2012). *CEFR-based framework for ELT in Japan*. Available from: http://www.cefr-j.org/english/index-e.html

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman Scale and parametric item response theory. *Political Analysis*, *11*, 139-163.

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*, 281-300. http://dx.doi.org/10.1191/0265532205lt309oa

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal, 91*, 676-679. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_9.x