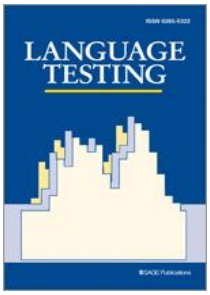
Self-Assessment of Japanese as a Second Language: The Role of Experiences in the Naturalistic Acquisition



Yuichi Suzuki
University of Maryland, USA

Abstract

Self-assessment has been used to assess second language proficiency; however, as sources of measurement errors vary, they may threaten the validity and reliability of the tools. The present paper investigated the role of experiences in using Japanese as a second language in the naturalistic acquisition context on accuracy of the self-assessment. Results revealed that experiential factors played a significant role in the measurement errors introduced by the self-assessment. The asymmetry pattern emerged, whereby less experienced second language speakers appeared to overestimate their ability, whereas those with more experience underestimated their language skills. The Rasch analysis identified **poor**-fit items in the self-assessment survey, and the subsequent qualitative analysis of the items indicated that the greater misalignment was related to the items including the more difficult tasks with which participants had relatively little experience. Implications for development and use of self-assessment are thus discussed in relation to experiential factors in self-assessment.

I Introduction

Self-Assessment (SA) has become popular as an alternative and supplementary tool for assessing the second language (L2) speakers' ability. Although SA has been applied to classroom instructions to motivate and improve learning in the framework of self-reflective practices (e.g., Leger, 2009 and Oscarson, 2013), the present paper focuses on the issue of the validity of SA as a proficiency indicator. Extant empirical evidence suggests that SA can be employed effectively to estimate L2 proficiency in some contexts, such as placement purposes (Blanche & Merino, 1989; Luoma, 2012; Ross, 1998 for review); however, the different sources of variability seem to influence the accuracy of self-ratings in SA. Among the factors affecting the accuracy of SA, the present study particularly examines the role of L2 experiences, which have been found to be one of the major factors in classroom settings (Butler & Lee, 2006; Ross, 1998). The work presented here further investigates whether this experiential factor induced measurement errors in SA for advanced Japanese L2 speakers after varying periods of residing in Japan.

II Literature Review: Self-assessment as a Subjective L2 Proficiency Indicator

Many researchers have examined the correlations between self-assessments and criterion measures, such as teachers' ratings, final grades, or objective tests. The results of these studies generally support the validity of self-assessments as a tool for assessing L2 proficiency levels. Findings of a meta-analysis conducted by Ross (1998) indicated that the magnitudes of relationships between SA and the criterion variables for four different skills (i.e., reading, listening, speaking, and writing) were robust, whereby the average coefficients ranged between $r = .524$ and $r = .650$. There seems to exist, however, a large variation or measurement errors specific to SA. Sources of measurement errors have been explored in previous studies on this issue, which revealed three major sources of variability in the accuracy of self-assessments, namely item characteristics (Heilenman, 1990; Ross, 1998), learners' individual characteristics (AlFallay, 2004; Anderson, 1982; MacIntyre, Noels, & Clément, 1997), and skill types (Ross, 1998).

A meta-analysis of L2 SA revealed that the higher correlations have been reported in receptive skills (i.e., listening and reading) compared to those pertaining to productive skills (i.e., speaking and writing) (Ross, 1998). With respect to item characteristics, the wording of the test items, as well as the skill domain, leads to the bias in the responses (Heilenman, 1990), as it appears that most individuals find it easier to respond to positive items (e.g., My spoken French is generally quite correct.), compared to negatively worded

items (e.g., My spoken French contains many mistakes.). Thus, this disparity creates response biases.

Researchers have investigated the effects of a wide range of individual characteristics on SA, including anxiety (MacIntyre et al., 1997), attitude/personality factors (Butler & Lee, 2006), self-esteem (AlFallay, 2004; Anderson, 1982; Wesche, Morrison, Ready, & Pawley, 1990), instrumental motivation (AlFallay, 2004), L2 proficiency (Heilenman, 1990), and experiences (Butler & Lee, 2006; Ross, 1998). These variables have been found to influence the accuracy of SA, and the results provide implications regarding the type of variables that should be controlled for in SA. Among those, the present study takes up the role of experiences because it is reasonably assumed that the amount of actual experience should be directly related to the accuracy of the responses in SA. The previous research particularly focused on the experience in classroom settings, and the current study aimed at extending the scope of experiential factors to the natural acquisition setting. The following review critically evaluates the existing studies that investigated the effects of experiences on the tasks in SA.

Following his meta-analysis of self-assessment in L2, which revealed significant variation in the SA accuracy, Ross (1998) attempted to empirically examine the influence of experience. In his study, instructed L2 English learners took part in listening comprehension activities, after which they took two types of SA. One format of the SA was closely matched to the contents of the instruction (experienced), whereas the other format was different from that used during classroom activities (abstract). Ross predicted that accuracy of SA would be higher when the criterion was achievement-based (experienced), rather than proficiency-based (abstract). The results supported the initial hypothesis, in that objective L2 proficiency test scores were predicted better by the SA on the experienced items, compared to the SA on the items that were not.

Similarly, Butler and Lee (2006) examined the validity of SA in a study in which elementary school EFL students in South Korea participated. Their findings revealed that the self-rating accuracy improved, in light of the concurrent validity of an objective L2 proficiency measure and teachers' assessment, when the SA was administered after the task was completed (i.e., experienced), compared to when the SA was administered without any context (i.e., abstract).

In sum, findings of the two empirical studies (Ross, 1998; Butler & Lee, 2006) suggest that actual experiences in the tasks addressed in the SA questionnaire influence the self-rating accuracy. These studies followed well-controlled experimental research protocol in which the experiential factors were examined via the comparison between the tasks experienced in classroom and the decontextualized tasks. However, there is an

evident paucity of research studies that have investigated experiential factors in self-assessment in more naturalistic L2 learning contexts, as SA research tends to focus on classroom contexts. In other words, no research has investigated the role of experience from real-life target language use on the accuracy of SA. There are growing interests in assessing what L2 learners can do in the world beyond the classroom setting (Little, 2006), and examining the validity of SA will provide a new insight on this issue when administering SA surveys to L2 speakers who are acquiring the target language in naturalistic acquisition context. The present study thus aims to fill this gap in the extant knowledge by examining the influence of experiences of actual target-language use during residence in a target language-speaking context.

III Interpretive Validity Arguments

In order to examine the validity of the Can-do Statements (CDSs) in relation to the role of experiences, the present study used the argument-based approach proposed by Kane (1992, 2004). A network of interpretive arguments was laid out for content validity of the CDSs (Kane, 2006). To build a validity argument for the target domain, the CDSs survey contained an assortment of linguistic activities that ranged from very commonly experienced activities to some possibly less commonlyexperienced activities. As experiences have been found to play a large role in the validity of SA (Butler & Lee, 2006; Ross, 1998), the present study particularly focused on the role of experiences in validation processes. The CDSs questionnaire was developed mainly for college students in Japan, and tasks that are supposedly encountered by them frequently were chosen (see the Method section). The underlying assumption is that activities in the CDSs should be well experienced by test-takers so that they can accurately assess what they can/cannot do.

In order to back up the validity argument, three sets of warrants were proposed. First, if the CDSs reflects the L2 ability properly, then there is a close relationship between the CDSs and object proficiency tests (Warrant 1). The length of exposure to L2 does not affect the accuracy of responses in the CDSs survey (Warrant 2). Additionally, a high internal consistency of the SA would be observed if the CDSs survey is not influenced by length of exposure and contact with the target language in the target-language community (Warrant 3).

The rebuttal for the first warrant would be a weak or no relationship between the CDSs and objective proficiency tests. For the second and third warrants, if the length of exposure to L2 influences the accuracy of responses in the CDSs survey, it would serve as rebuttal to the validity claim.

IV Method

1 Participants

Sixty-three L2 Japanese speakers whose first language (L1) is Chinese participated in the study (11 male and 52 female). They were all advanced L2 speakers, and the requirement for participation in the study was advanced Japanese proficiency equivalent to the most-advanced (N1) and the second-advanced (N2) in the standardized Japanese Language Proficiency Test (JLPT) (https://www.jlpt.jp/e/index.html). They were recruited through fliers distributed in college, announcements in classes, and word of mouth. JLPT N1 and N2 are roughly equivalent to ACTFL Superior and advanced on the OPI scale, respectively (Kanno, Hasegawa, Ikeda, & Ito, 2005).

All participants were late L2 speakers, who had arrived in Japan after the age of 18 (Table 1). With the exception of one speaker, all participants received classroom instruction before and/or after they came to Japan. Starting age of instruction (AOI) was as early as 14, and the mean length of instruction (LOI) was 39.03 months. Many of the participants had majored in Japanese, and received four years of instruction at a university in China. The majority of the participants were university students or academic scholars, with 16 undergraduate students, 12 research students, 26 master's students, two PhD students, one post-doctoral scholar, one visiting scholar, one vocational student, and two who were office workers. Their mean age was 24.65 at the time of testing.

In order to estimate the participants' L2 experiences, the questionnaire included items pertaining to the length of residence in Japan and the actual amount of experience in Japanese during the stay. The mean length of residence (LOR) was 26.76 months (range 3-158). Since the CDSs in the present study focused on the reading skills, the participants were asked to indicate the number of hours per day they typically devoted to reading Japanese during their stay (e.g., the Internet, TV, books, music, news, comic books, etc.). Thus, in the subsequent data analysis, cumulative reading hours (Reading Experiences, RE) were calculated as the product of the number of reading hours per day and number of years of stay. The average RE was 8.27 hours. The LOR (number of years of stay) and RE (cumulative reading hours) were used as indicators of learners' experiences in the present study.

Table 1

Background Information of L2 Speakers

	Age	Age of Instruction	Length of Instruction	Age of Arrival	Length of Residence	Reading Experiences
Mean	24.65	19.13	39.03	22.17	26.76	8.27
SD	4.34	2.20	20.17	3.16	26.82	8.33
Min	20	14	0	18	3	0
Max	52	25	125	42	158	36

Note 1. Age is at the time of testing in years. Length of Instruction (in months), Age of Arrival (in years), Length of Residence (in months), Reading Experiences (in hours).

Note 2. RE = average reading hours per day * length of residence in years

2 Instruments

a Can-do statements. A self-assessment survey (can-do statements; CDSs) was administered, which was developed mainly for those individuals who have attained N1 and N2 in JLPT (Shimada, Saegusa, & Noguchi, 2006). The target-use domain in the CDSs questionnaires focused on college life, but it not only involved academic activities but also activities in life outside of universities. In order to develop the can-do statements, 173 descriptors of language behaviors were first selected by referring to the previous studies on foreigners' use of Japanese and contents from textbooks used in colleges and Japanese language schools. The 173 descriptors were further narrowed down to 60 descriptors (15 for each of the four skills) that are 1) needed to lead daily-life and academic activities in college, 2) activities supposedly experienced by the test-takers, and 3) concrete involving authentic situations.

In the questionnaire, the participants were required to indicate how much they could do in Japanese on a 7-point Likert scale, with 15 CDSs pertaining to the reading skills (Appendix A). All the selected items were positively worded. Since SA on reading skills has been found to be most accurate and highly correlated with objective L2 proficiency tests compared to other skills (Ross, 1998; Shimada et al., 2006), it was expected to be less influenced by L2 experiences. If the influence from experiences is identified even in the reading skills (i.e., the skill which can be assessed most reliably), it suggests that the effects of experience are robust. The reliability of the reading section of the questionnaire reported in Shimada et al. (2006) was high (Crobach's $\alpha = 0.953$).

b Japanese C-test. One of the two objective proficiency tests was a Japanese C-test (Shin, 1990). A C-test was constructed to measure Japanese L2 proficiency in the target population of this study (i.e., advanced Japanese L2 speakers). It is commonly used as a good objective measure of L2 proficiency (e.g., Eckes & Grotjahn, 2006), and the current study chose the C-test to measure vocabulary and grammar knowledge in written modality. One passage was selected for the development, which contains vocabulary at N1 and N2 levels (Appendix B). This text was chosen because the difficulty of the passage was appropriate for advanced L2 speakers. More importantly, the passage dealt with a common and familiar topic for everyone staying in Japan so that the topic specificity does not bias the test scores of the participants in the current study (see Appendix B). We used the Japanese readability scoring system alpha version to compute the difficulty of the passage (http://jreadability.net/). This scoring system classifies texts into six readability levels, and the readability score of the current passage was 2.19. This corresponds to the second most difficult text level in the readability scoring system, and those who can read the text at this level can understand most of the technical passages and complex structures in the literary art. The passage consists of seven sentences and the average length of the sentence was 32.71 characters.

In order to create blanks for the test, we first segmented the sentences into a *bunsetsu* or a clause. For example, a sentence like *ringo wo tabemasu* (Apple-OBJECT eat) consists of two clauses. The content word (i.e., *ringo*, apple) followed by the object-marking particle *o* forms a unit of *bunsetsu*. If a unit of content words is not accompanied by function words (i.e., *tabemasu*, eat), the content word forms another unit. Based on the standard C-Test development procedure (Klein-Braley, 1997), partial deletions started at the second word of the second sentence in a passage, from which every second word was partially deleted. For the unique characteristics of Japanese (see Lee-Ellis, 2009 for more detailed rationales for the development of C-tests for Korean, which is very similar to Japanese grammatical constructions), the partial deletion was made to the latter half of each content word, including function words (particles) and inflections. The first *bunsetsu* in the example above (i.e., *ringo wo*) would be deleted as follows: *rin*__ __.This procedure created blanks the respondents needed to fill, thus indicating their knowledge of both content words and particles. One-syllable words and proper nouns that could not be recovered from the context were left intact. When a content word had an odd number of syllables, half of the syllables minus 0.5 were left blank (e.g., 3 → 1, 5 → 2).

Japanese orthography consists of three types of scripts, *hiragana*, *katakana*, and *kanji* (Chinese characters). Japanese orthography is complex, and the same words are sometimes written in three different ways. For example, *tamago*, 'egg,' can be written in

three scripts, for instance (卵, たまご, タマゴ). Some words can be written in a hybrid of kanji and hiragana (e.g., 買う to buy), and inflectional morphemes (verb endings) are always written in *hiragana*. Given this complex feature of Japanese orthography, giving hints about the number of letters and types of scripts to be filled out would be beneficial for both test-takers and testers. In the construction of the current C-test, one type of blank (circle) was used for *hiragana* and *katakana*, and the other type of blank (square) was used for *kanji*. This narrowed down the possible answers, which made the questions and scoring easier and simpler (Shin, 1990).

Due to the specific deletion method of the current C-test (whereby everything after the second half of content words is deleted, including attached functional words), alternative answers that are different from the target answer were obtained from both NSs and L2 learners. These possible answers were examined by the investigator for their grammatical and contextual appropriateness, after which the answer key was created, and was used in the scoring procedure.

c Elicited Imitation. Another proficiency test included in the study was Elicited Imitation (EI). This test is controversial on exactly what construct it really measures, especially in the field of Second Language Acquisition. Of particular interest to SLA researchers is whether L2 competence consists of implicit knowledge, explicit knowledge, or a combination of both (Jessop, Suzuki, & Tomita, 2007), and they have attempted to validate EI as an implicit knowledge measure (Bowles, 2011; Ellis, 2005, 2006; Ellis et al., 2009; Erlam, 2006). Despite the different views on the constructs that EI taps, there is a consensus among L2 researchers that EI taps linguistic knowledge (Bley-Vroman & Chaudron, 1994). EI has also been found to be correlated highly with other proficiency measures (Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008; Henning, 1983). The current study used the EI task as a measure of L2 proficiency. It is noted that the modality of EI is aural, which is different from the C-test. The EI procedure consisted mainly of the following three components: (1) auditory stimulus sentence processing; (2) a comprehension question; and (3) imitation of the sentence.

The EI task included five different Japanese grammar structures that were known to be difficult to acquire for Chinese speakers: (1) transitive/intransitive verb pairs, (2) wa/ga in a adverbial clause, (3) the wa/ga in a relative clause, (4) the genitive marker *no*, and (5) locative particles *ni*/*de* (see Suzuki, 2013 for detailed information). Sixteen stimuli sentences were created for each of the five structures (k = 80 in total).

The responses were scored based on Erlam's (2006) criteria: (1) obligatory occasion created – supplied; (2) obligatory occasion created – not supplied; and (3) no obligatory occasion created (see Erlam, 2006 for the detailed procedure). Credit was given only to the

first category, while the remaining two categories were scored as incorrect. Giving no credit to the third category may need justification because structural modification does not necessarily mean that speakers intentionally avoid the structure. However, given that responses by NSs and L2 speakers who scored high in the EI task rarely belonged to the third category, it is likely that most of the responses in this category are due to the lack of ability to use those target structures. As the two sample sentences given in the instructions did not allow any structural modifications, while permitting word substitutions, structural modifications were not encouraged.

3. Procedure

Every participant engaged in two separate sessions that took place on different days. The interval between the two sessions was approximately one week for most of participants. In the first session, participants performed the EI. No feedback was provided after the EI task to make sure that the responses of the CDSs questionnaire were not affected by the results. In the second session, they first answered the CDS questionnaire before they took the C-test so that the performance of the reading test would have no influence on the responses in the CDSs survey.

4. Analysis

In order to provide evidence for Warrant 1, correlation coefficients between the CDSs and object proficiency tests were computed. A simple regression analysis was also conducted on the CDSs with the C-test and the EI task separately. The main focus of the study is to investigate whether the amount of L2 experience influences the accuracy of CDSs (Warrant 2). In order to test this, the CDSs were regressed on the objective tests to compute the residual (expected score – observed score). This residual or discrepancy between the SA and the objective tests was compared with the L2 experiences (LOR and RE). Lastly, the data were further analyzed to examine the third warrant—internal consistency of the CDSs survey. We computed outfit statistics based on the Rasch model. If some poor-fit items were identified by the Rasch analysis, the content of the items would be closely inspected to examine the effects of experience.

V Results

1 Descriptive Statistics

Descriptive Statistics for CDSs, the C-test and the EI task are presented in Table 2. The mean score achieved on CDSs was 5.99 out of 7, indicating that respondents were confident in most of the tasks asked in the CDSs. For the C-test, the mean was 19.06, and the distribution of the data suggests that the assessment was able to discriminate advanced L2 speakers of Japanese. The mean EI score of 58.11 out of 80 was achieved by L2 speakers. Reliability indexed by Cronbach's alpha was high in all the tests: the CDSs (α = .91), the C-test (α = .81), and the EI task (α = .94).

Table 2

Descriptive Statistics for EI, C-test and CDSs

	Mean	SD	Min	Max	Skewness	Kurtosis	Possible Max	α
CDSs	5.99	0.60	3.93	7.00	-1.06	1.83	7	.91
C-test	19.06	4.78	9	28	-.30	-.87	30	.81
EI	58.11	14.18	15.00	76.00	-0.97	0.53	80	.94

In order to examine the concurrent validity of the two objective tests, the Pearson's correlation coefficient between the C-test and the EI task was computed. A strong positive relationship was detected ($r = .757$, $p < .001$), providing evidence that the two objective tests covary with each other.

Next, the relationships of the L2 experiences with these tests were explored (Table 3). The correlations of L2 experiences and reading experiences with CDSs were significantly positive, and the correlation coefficient was slightly higher with reading experience. The C-test also correlated significantly with reading experiences, but not with the LOR. The correlations of the EI task with LOR and reading experiences were not meaningful or significant ($r = .072$, $r = .195$, $p > .1$). It seems that actual reading experiences seem to provide a more sensitive measure of L2 experiences than the overall estimate of L2 use indexed by LOR in the present study. The correlation between LOR and reading experiences was high ($r = .736$, $p < .01$), but not identical. Thus, both experience variables were used in the further analysis.

More experience or exposure to written texts is likely to lead to higher ratings of CDSs and L2 proficiency measured with the C-test, which used the written passage. It adds further support for the validity of the CDSs and the C-test in that both measurements tap the same construct (i.e., L2 ability to read written texts) to some extent. The central question addressed in this paper is whether the *inaccuracy* in self-ratings of the CDSs can be explained by the experience, which will be discussed in the next section.

Table 3

Correlations between L2 Experiences, CDSs, C-test, and EI

	LOR	RE	CDSs	C-test	EI
LOR	-	.736**	.376**	.145	.072
RE		-	.415**	.272*	.195
CDSs			-	.251	.210
C-test				-	.757**
EI					-

Note. LOR=Length of Residence, RE = Reading Experiences.

2 Influence of L2 Experiences on Self-Assessment

In order to examine the discrepancy between the CDSs (i.e., perceived ability) and the objective L2 proficiency tests (i.e., actual ability), the CDSs were regressed on the objective tests to compute the residual (expected score – observed score). If the expected score from the objective L2 proficiency test score is equal to the (observed) CDSs score, the residual will be zero. In other words, a residual score of zero indicates the most accurate self-rating on what they can do in the CDSs. Negative residuals indicate overestimation of their can-dos because the observed score (i.e., self-ratings) is higher than the predicted value (by objective proficiency scores). Similarly, positive residuals indicate underestimation of the ability, as the observed self-rating score is lower than the predicted score. Since the C-test and the EI task are highly correlated to each other ($r = .757$, $p < .01$), the multiple regression analysis with C-test and EI as predictors on the CDSs was not conducted to avoid attenuating the predicting powers (i.e., collinearity). Instead, a simple regression analysis was conducted on the CDSs with the C-test and the EI task separately.

A simple regression analysis was conducted on the CDSs from the C-test to compute the residual (expected scores – observed scores). The analysis revealed that the model was almost statistically significant, $R^2 = .063$ (adjusted $R^2 = .047$), $F(1, 59) =$

3.977, p = .051. The unstandardized and standardized regression equations are presented in Table 4.

Table 4

Regression on CDSs with as C-test a Predictor

	Unstandardized Coefficients		Standardized Coefficients		
	B	*SE*	*β*	*t*	*p*
(Constant)	5.36	0.32	0.00	16.61	0.00
C-test	0.03	0.02	0.25	1.99	0.051

As in the previous analysis, a simple regression analysis was conducted on the CDSs from the EI task to compute the residual. The findings revealed that the model was almost as good as the previous model, R^2 =.044 (adjusted R^2 = .028), $F(1, 59) = 2.71$, p = .105. The unstandardized and standardized regression equations are presented in Table 5.

Table 5

Regression on CDSs with as EI a Predictor

	Unstandardized Coefficients		Standardized Coefficients		
	B	*SE*	*β*	*t*	*p*
(Constant)	5.46	0.33	0.00	16.60	0.00
EI	0.01	0.01	0.21	1.65	0.11

The two regression analyses revealed weak relationships between CDSs and the two objective measures, indicating that the influence of L2 experiences should be closely examined in order to ascertain whether it contributed to the measurement errors in the CDSs.

In order to examine the effects of L2 experiences on the residuals or discrepancy between the CDSs and L2 proficiency measured with the objective test, the correlations between the residuals (from the C-test and the EI) and L2 experiences (LOR and reading

experiences) were computed. Significant positive relationships were revealed: residual with C-test * LOR (r = .351, p < .01), residual with C-test * reading experiences (r = .358, p < .01), residual with EI * LOR (r = .369, p < .01), and residual with EI * reading experiences (r = .382, p < .01). Figures 1 and 2 show the scatter plots between residual and L2 experiences, revealing that overestimation (negative residuals) was observed in the less-experienced L2 speakers as well as underestimation. Although the number of data points corresponding to the more experienced speakers is lower, a few of these participants overestimated their can-dos. The locally weighted scatterplot smoother (LOWESS) lines were superimposed to demonstrate a line of best fit for the data. The lines show a decline in participants with less experience, but the line gets smoother as experience accumulates. The dip(s) of the LOWESS lines in less-experienced L2 speakers supports the idea that L2 speakers with less experience tended to overestimate their ability, although the magnitudes of the decline were not large. Residuals pertaining to experienced speakers tended to be positive, that is, they tend to underestimate what they can do and to be more conservative in assessing their L2 ability.

These results support the premise that inaccurate self-ratings in CDSs can be partly explained by the L2 experiences. L2 speakers with less experience have difficulty estimating what they can do (indicated by higher variability in the residuals), but with more experience, their self-ratings will be more accurate. Furthermore, experienced speakers tend to underestimate their ability, most likely because they are more keenly aware of their own limitations. In other words, they learned what they could not do through extensive experiences, which is why they do not overestimate their ability. In contrast, L2 speakers with less experience tend to overestimate what they can do more than experienced speakers, due to lack of the actual experiences referred to in the CDSs. Overestimation has been found to be more evident in less proficient learners than in more proficient learners in the foreign language classroom setting (Heilenman, 1990). Even though the present study focused on the *advanced* L2 speakers who are in the natural acquisition setting (and are less likely to overestimate), the amount of experience during the residence in Japan seems to exert influence on the accuracy of SA.

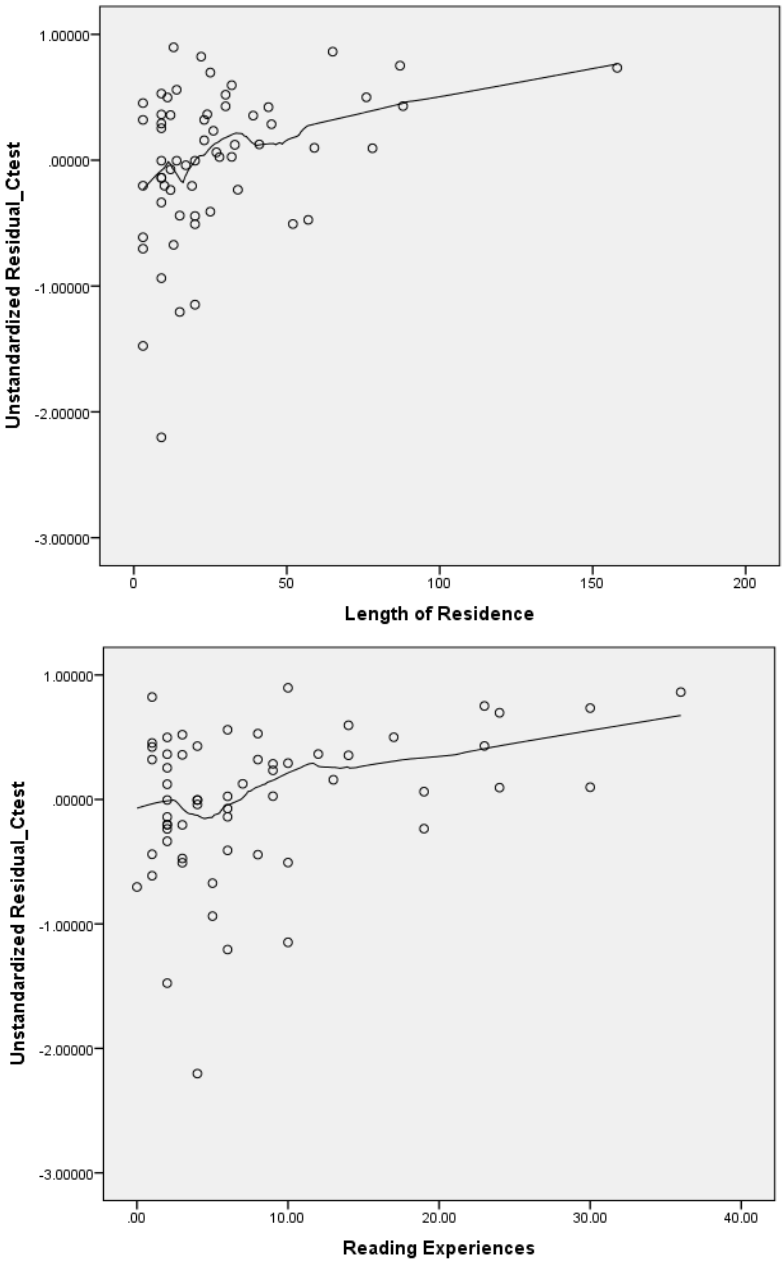


Figure 1. Scatter Plots between L2 Experiences and Residuals from C-test

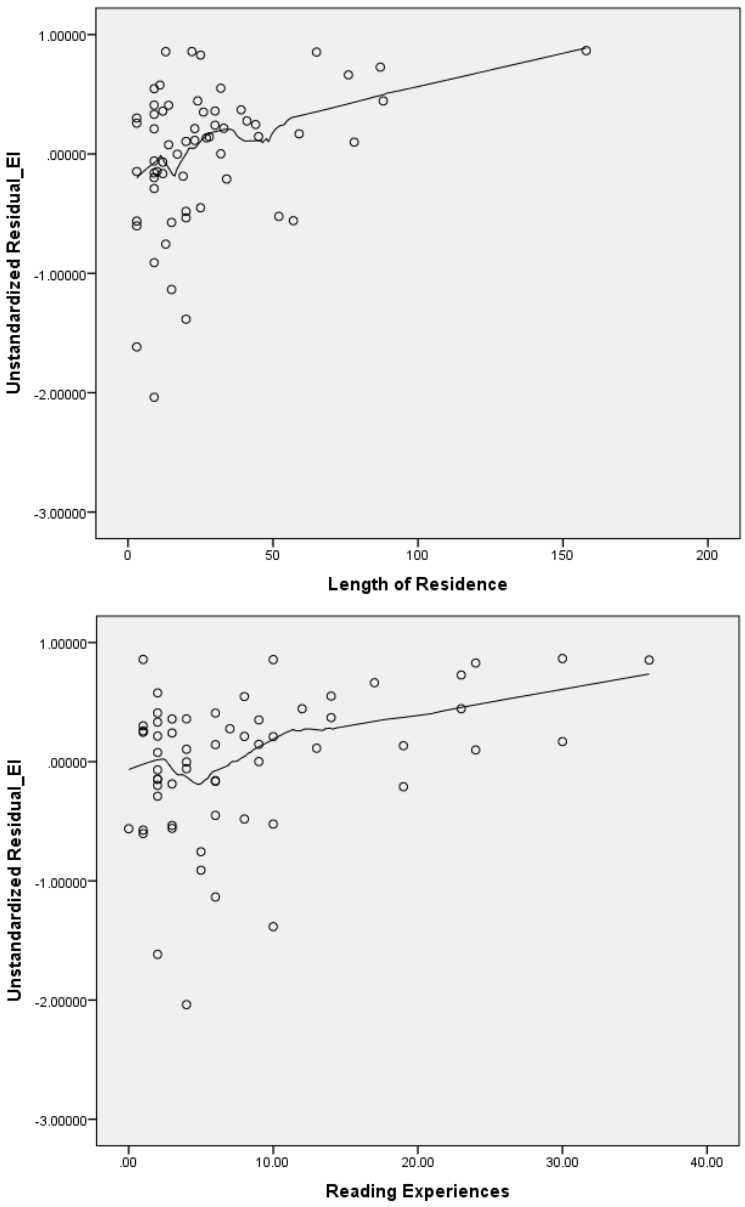


Figure 2. Scatter Plots between L2 Experiences and Residuals from Elicited Imitation

3 Content Analysis

In order to examine the source of unreliable variations in CDSs, their characteristics were inspected more closely using the Rasch Rating Scale model (Bond & Fox, 2007). The Rasch analysis was conducted to identify the items that L2 speakers answered inconsistently (misaligned items, referred to as misfit) and the ones that received more consistent responses (aligned items, referred to as fit).

Fit statistics in Rasch analysis generates two types of statistics, infit and outfit. Unstandardized outfit statistics were used as an indicator of (mis-)fit. Outfit was chosen because the CDSs survey has a small number of items. "Infit is more sensitive to the pattern of responses targeted on the person, and vice-versa…Outfit means outlier-sensitive fit. This is more sensitive to responses to items with difficulty far from a person, and vice-versa." (Linacre, 2012, p.293) With the small number of items in the CDSs survey, it is more logical to use outfit because it is more stable than infit because outfit is more sensitive to items far from a person. The mean square of the outfit is the average value of the squared residuals for each item. The residuals represent the differences between the expected values from the Rasch model's theory and the observed values. In other words, the outfit values index the unexpected or unreliable behavior of items. The Rasch analysis also provides estimates for item difficulty with standard errors on the logit scale, whereby items with higher positive values are more difficult, and the more negative values indicate less difficulty.

The three items with worst outfit are presented in Table 6: item 5 (Can you read and understand novels?), item 1 (Can you read and understand newspaper editorials?), and item 9 (Can you read and understand the questionnaire given before having an examination at a medical office or hospital?). Apparently, all the tasks in those items are less likely to have been encountered by L2 speakers (as well as native speakers), and experiences in the tasks may have varied greatly among them. In addition to fewer experiences, the tasks in item 5 and item 1 are more abstract, and it is harder to give self-ratings based on one's own episodic memory. Although the task in item 9 is more specific and concrete, most of the L2 speakers, especially those that have spent only limited time in Japan, have never had an examination at a hospital during their stay. The mean for difficulty estimate was set at 0 (SD = 1.04). Given this, all three items were relatively difficult. Among the 15 items, item 5 was the most difficult (1.93), item 1 the third (1.38), and item 5 the fifth (0.85).

Table 6

Worst Three Fitted Items according to Outfit Statistics

	Difficulty Estimate	S.E.	Outfit (Mean Square)	Outfit (Standardized)	Exact Observed %	Match Expected %
Item 5	1.93	0.17	1.24	1.3	45.8	49.2
Item 1	1.38	0.17	1.33	1.7	44.1	48.1
Item 9	0.85	0.18	1.71	3.3	42.4	50.3

In sum, the unstable variations in the CDSs largely stem from experiences (and difficulty) in the tasks. This further corroborates the positive relationship between residuals and L2 experiences.

VI Conclusions

The present paper examined the validity of the Can-do Statements (CDSs) with a particular focus on the role of experiences in the natural acquisition setting. To test the content validity argument, three sets of warrants were proposed and examined. The relationship between CDSs and two objective L2 proficiency tests was not strong (Warrant 1), which suggested that some other factors are contributing to this weak relationship. The relationship between CDSs and the two objective tests was weaker than the results obtained in other studies (see meta-analysis in Ross, 1998). This may be due to the fact that participants in the present study are not classroom learners and have more variability in how they learned Japanese as a second language. The participants were advanced L2 speakers, and they might have achieved that level through a variety of experiences. For instance, they are confident with the statements in the CDSs because they encounter the specific situations many times in real life. It may be possible that although some of them cannot do well on the objective tests, they are more confident to perform the actual real-life tasks represented in the CDSs, which could have attenuated the strength of relationship between the objective tests and the CDSs.

The current study further examined whether the length of residence and the amount of exposure to reading materials influence the accuracy of responses in the CDSs survey (Warrant 2). The residual analysis demonstrated that the inaccuracy in self-assessment could be partially explained by the experiential factors.

This suggests that experiential factors play a significant role in accuracy in the CDSs for advanced L2 speakers who learned the language in the immersion context (i.e., living in Japan). This finding is consistent with those reported in previous studies that

found that experiential factors improve the accuracy of SA in a classroom context (Butler & Lee, 2006; Ross, 1998). These consistent results across experimental and naturalistic contexts provide further evidence regarding the impact of experiential factors on SA. Furthermore, there was a reverse tendency, whereby L2 speakers with less experience tended to overestimate their reading skills in CDSs more, whereas those with more experience underestimated their skills.

The results from the Rasch analysis identified poor-fit items in the CDSs. The qualitative analysis of the tasks in the CDSs suggested that the higher misfit was identified in less frequently experienced (more difficult) items. This finding further supports the significant influence of experiences on the accuracy of SA (Warrant 3).

Overall, the present study provided the rebuttals for all three warrants for the CDSs, and highlighted the role of experiential factors in SA. It underscores the challenges in developing valid CDSs and calls for more careful development and use of CDSs particularly in the naturalistic acquisition context. The findings obtained in the study have implications for the development and use of CDSs. When choosing the tasks in the can-do statements, it should be ensured that the target L2 speakers have had sufficient prior experience with those tasks. Efforts should be made to avoid asking for self-ratings for the tasks that are not relevant to the target population. It may be appropriate to directly ask test-takers to rate their prior experiences in the tasks involved, so that the items corresponding to the tasks in which the test-takers have no prior experience may be excluded in order to reduce the errors. Furthermore, the fact that more experienced speakers tended to underestimate their perceived ability, and less experienced typically overestimated it, even among advanced learners, is problematic. A more careful consideration should be given when administering SA surveys to L2 populations with different L2 experiences.

That being said, the current study opens several avenues for future research. First, the objective L2 proficiency measures employed in the study (i.e., C-test and EI) were decontextualized tests, and a fairer comparison can be made between CDSs and the objective tests if the latter require the performance of the actual task asked in the CDs.

Second, the skills the CDSs addressed were only related to reading skills. Other skills may be more influenced by experiential factors because (1) L2 speakers usually start learning second languages in classrooms, where instructions place more emphasis on reading (i.e., more experiences in reading), and (2) it has been reported that other productive skills, such as listening and speaking, are harder to assess in SA.

Third, previous studies have examined a wide range of individual characteristics on SA (e.g., anxiety, motivation, attitudes), which were beyond the focus of the present study.

Given a paucity of research studies that have investigated SA in naturalistic L2 learning contexts, future research should examine how these factors influence SA.

Last, participants with different backgrounds may produce different patterns in SA, allowing different relationships between SA accuracy and experience to be revealed. In this study, Chinese was the first language of all the participants, most of whom were university students. Although the current CDSs were created mainly for university students, more test-takers with different L1 backgrounds and different statuses (e.g., office workers) should be recruited to evaluate the influence of experience in SA. Additionally, Chinese speakers have an advantage in being familiar with kanji or Chinese characters. Further research is needed to generalize the present findings for a wider population.

VII References

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self-and peer-assessment. *System, 32*(3), 407-425.

Anderson, P.L. (1982). Self-Esteem in the Foreign Language: A Preliminary Investigation. *Foreign Language Annals, 15*(2), 109-114.

Blanche, P., & Merino, B.J. (1989). Self-Assessment of Foreign-Language Skills: Implications for Teachers and Researchers. *Language Learning, 39*(3), 313-338. doi: 10.1111/j.1467-1770.1989.tb00595.x

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 245-261).

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*: Routledge.

Bowles, M.A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 33*(2), 247-271.

Butler, Y.G., & Lee, J. (2006). On-Task Versus Off-Task Self-Assessments Among Korean Elementary School Students Studying English. *The Modern Language Journal, 90*(4), 506-518. doi: 10.1111/j.1540-4781.2006.00463.x

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290-325.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition, 27*(2), 141-172.

Ellis, R. (2006). Modelling Learning Difficulty and Second Language Proficiency: The Differential Contributions of Implicit and Explicit Knowledge. *Applied Linguistics, 27*(3), 431-463.

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Tonawanda, NY: Multilingual Matters.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464-491.

Graham, C.R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). *Elicited imitation as an oral proficiency measure with ASR scoring.* Paper presented at the Proceedings of the 6th International Conference on Language Resources and Evaluation.

Heilenman, L.K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*(2), 174-201.

Henning, G. (1983). Oral Proficiency Testing: Comparative Validities of Interview, Imitation, and Completion Methods. *Language Learning, 33*(3), 315-332. doi: 10.1111/j.1467-1770.1983.tb00544.x

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review/La Revue canadienne des langues vivantes, 64*(1), 215-238.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological bulletin, 112*(3), 527.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170.

Kane, M.T. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131-153). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Kanno, K., Hasegawa, T., Ikeda, K., & Ito, Y. (2005). *Linguistic Profiles of Heritage Bilingual Learners of Japanese*. Paper presented at the Proceedings of the 4th International Symposium on Bilingualism.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14*(1), 47-84.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing, 26*(2), 245-274. doi: 10.1177/0265532208101007

Linacre, J.M. (2012). A user's guide to WINSTEPS MINISTEP: Rasch-model computer programs. from http://www.winsteps.com/a/winsteps-manual.pdf

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*(3), 167-190.

Luoma, S. (2012). Self-assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 5169–5174). New York, NY: Wiley-Blackwell.

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*(2), 265-287.

Oscarson, M. (2013). Self-assessment in the classroom. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. II: Approaches and Development, Part 6. Assessment and Learning, pp. 712–729). New York: Wiley-Blackwell.

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

Shimada, M., Saegusa, R., & Noguchi, H. (2006). Nihongo Can-do-statements wo riyoushita gengo koudou kijutsu no kokoromi: Nihongonouryokujukensha wo

taishou to shite. [Attempt of Language Behavior Description by Using Japanese Can-do-statements: Testees of Japanese Language Proficiency Test]. *Sekai no nihongokyouiku, 16*, 75-88.

Shin, K. (1990). From a Japanese Cloze Test to a Japanese Modified C-Test : Concerning Scoring Ploblems. *Studies in language and culture, 11*(2), 213-225.

Suzuki, Y. (2013). *Deconstructing Elicited Imitation: Evidence from the Word-Monitoring Task.* (Unpublished Qualifying Paper), University of Maryland, College Park.

Wesche, M., Morrison, F., Ready, D., & Pawley, C. (1990). French Immersion: Postsecondary Consequences for Individuals and Universities. *Canadian Modern Language Review, 46*(3), 430-451.

Appendix A. Can-do Statements

Please assess your current ability to use Japanese in the following situations. For each question, please circle the number between 1 and 7 that you believe best indicates your level. Mark the circle on the intersection point on the graph below the number, not in between the numbers. If you have not experienced a given situation, please try to imagine the situation and choose the number that best applies. 1-7 can be interpreted in the following manner:

1 2 3 4 5 6 7

Cannot do at all — Can do to a certain extent — Can do with no problems

1. Can you read and understand newspaper editorials?

2. Can you read and understand posters, notices, and other printed materials posted around school?

3. Can you read and understand school rules and regulations?

4. When you look at the spines of books on the shelves in the library, can you find the book you are looking for?

5. Can you read and understand novels?

6. Can you read and understand the flyers in train stations, travel agencies, etc.?

7. Can you read and understand books, academic papers, etc., that are necessary for your studies?

8. Can you understand advertisements in trains, buses, etc.?

9. Can you read and understand the questionnaire given before having an examination at a medical office or hospital?

10. Can you read and understand things that are handwritten on blackboards, bulletin boards, etc.?

11. Can you read and understand newspaper articles about societal issues (incidents, accidents, etc.)?

12. Can you read and understand the required information on water, electricity, and gas bills?

13. Can you understand computer and machinery operating manuals?

14. Can you understand the notices and information sent by school or city hall?

15 Can you read and understand job search information (job advertisements, part-time work information, etc.)?

Appendix B. C-tests

文章を読んで、◯と□に単語を書いて、文章を完成させて下さい。◯には、ひらがなかカタカナが入ります。□には、漢字が入ります。
(Please complete the texts by filling out the circles and squares with Hiragana (Katakana) and Kanji, respectively.)

平成 24 年 3 月に、日本は日本食をユネスコの無形文化遺産として登録できるように申請を行いました。一口に「日□◯[1]食文化」といっても、郷土料理のような地□◯[2]文化、マナ◯◯[3]おもてなしの心◯◯[4]、様々な側□◯[5]あります。どのように登□◯[6]行うかに関◯◯[7]、専門家を中□◯◯◯[8]グループによって議□◯[9]重ねられてきま◯◯[10]。

日本の国□◯[11]南北に長◯[12]、海、山、里と豊か◯[13]自然が広が◯◯[14]いるため、各□◯[15]地域に根差◯◯[16]食材が利用さ◯◯[17]います。また、季□◯[18]移り変わりが明確◯◯◯[19]、四季折々の食□◯[20]利用されます。これらの多□◯[21]素材の味わ◯◯[22]活かす調理□□◯[23]調理道具も発達◯◯[24]います。一汁三菜を基□◯[25]する極め◯[26]健康的な和□◯[27]栄養バラ◯◯◯[28]理想的で肥満□□[29]や長寿にも役立◯◯[30]います。

A Short Summary in English

In order to designate Japanese food as an intangible cultural heritage, Japan applied to UNESCO in March, Heisei 24th (2012). A group of specialists has discussed how to pitch the argument for registration of Japanese food. Because Japan is full of rich nature and the continent extends widely from north to south, there are ingredients that originated from a diverse area. Four distinct seasons make seasonal ingredients available. Additionally, Japanese people developed cooking methods and tools for the variety of ingredients. It is said that well-balanced Japanese food contributes to longevity and the prevention of obesity.

Answer Keys

1. 本の
2. 域の、元の、方の
3. ーや
4. など
5. 面が
6. 録を

7. して
8. 心とした、心とする、核とする
9. 論が
10. した
11. 土は
12. く
13. な
14. って
15. 地で、地の、々の、県や
16. した
17. れて
18. 節の
19. なため、であり、なので
20. 材が
21. 様な、彩な、種の、数の
22. いを
23. 技術や、方法や、手順や
24. して
25. 本と、調と、本に
26. て
27. 食の、食は
28. ンスが、ンスは、ンスも
29. 防止、予防、改善、対策、解消、抑制、減少、回避
30. って