ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 4, No. 5, pp. 900-913, September 2013

doi:10.4304/jltr.4.5.900-913

Vocabulary Knowledge and Speaking Proficiency among Second Language Learners from Novice to Intermediate Levels

Rie Koizumi
Juntendo University, Japan

Yo In'nami
Shibaura Institute of Technology, Japan

***Abstract*—To remedy the paucity of studies on the relationship between vocabulary knowledge and speaking proficiency, we examine the degree to which second language (L2) speaking proficiency can be predicted by the size, depth, and speed of L2 vocabulary among novice to intermediate Japanese learners of English. Studies 1 and 2 administered vocabulary tests and a speaking test to 224 and 87 L2 learners, respectively. Analyses using structural equation modeling demonstrated that a substantial proportion of variance in speaking proficiency can be explained by vocabulary knowledge, size, depth, and speed. These results suggest the centrality of vocabulary knowledge to speaking proficiency.**

***Index Terms*—vocabulary size, depth, speed, L2 speech production, fluency, accuracy, syntactic complexity**

I. Introduction

Vocabulary has long been recognized as a vital component and a good indicator of second language (L2) performance and proficiency (e.g., Schmitt, 2010; Stæhr, 2009). However, compared to the numerous studies on associations between L2 vocabulary and reading (e.g., Qian, 2002; Van Gelderen et al., 2004), little research has been conducted into the relationships between L2 vocabulary and other L2 skills (Stæhr, 2009). Examples include Stæhr (2009) for listening, Schoonen et al. (2003) for writing, and De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012) for speaking. The current article focuses on the relationship between L2 vocabulary knowledge and L2 speaking proficiency among novice- to intermediate-level Japanese learners of English, by conducting two studies that use structural equation modeling (SEM).

A. Vocabulary Knowledge and Its Predictive Power

While researchers generally agree with regard to the multicomponential nature of vocabulary knowledge, various proposals have been put forward regarding what exactly constitutes vocabulary knowledge (e.g., Meara, 2005; Schmitt, 2010). One classification frequently employed involves the size and depth of vocabulary (e.g., Qian, 2002). Size, or breadth, expresses a quantitative dimension involving knowledge of a word form and a primary meaning, also described as the form-meaning link. Depth represents a qualitative dimension, defined as "how well a learner knows individual words or how well words are organized in the learner's mental lexicon" (Stæhr, 2009, p. 579), and includes knowledge of partial to precise meaning, word frequency, affix knowledge, syntactic characteristics, and lexical network.

In addition to size and depth, another lexical aspect that has recently attracted attention and been incorporated into vocabulary frameworks is speed of processing, or how fast learners can recognize and retrieve knowledge stored in the mental lexicon (e.g., Meara, 2005). Processing speed (often referred to as automaticity, efficiency, or fluency) of lexical access and retrieval is considered to play a crucial role in the use of vocabulary in real-life situations, as well as in L2 proficiency (e.g., Van Moere, 2012). This may be true especially of listening and speaking, which require on-line processing (Schmitt, 2010).

Of these multidimensional lexical aspects, size has been considered primary, because of the importance of the form-meaning link for vocabulary use (e.g., Laufer, Elder, Hill, & Congdon, 2004; Schmitt, 2010). A number of empirical studies have been conducted to examine the relative importance of size versus depth and speed in terms of predictive powers of L2 skills. Qian and Schedl (2004) investigated vocabulary knowledge and reading comprehension among 207 L2 learners of English at intermediate and advanced levels, and reported that 57% of variance of L2 reading scores was explained by size, with an additional 4% of variance explained by depth. A similarly large variance (54%) predicted solely by size was indicated by Qian (2002), with an additional 13% explained by depth (n = 217). Finally, Stæhr (2009) provided further support for these results, showing that 49% of L2 listening variance was accounted for by size, but just 2% by depth (n = 115). In sum, previous studies suggest that size can predict much of reading and listening variance, while depth contributes relatively little.

It should be noted that the proportion of variance explained by variables changes, depending on the order in which

independent variables are entered into the regression equation. The results from the studies described above were derived when size was entered first, followed by depth. The effect of this is that depth is able to predict only the remaining variance, that is, whatever was not predicted by size. Since size was highly correlated with depth, sharing a large variance with it (49% in Qian, 2002, r = .70; 71% in Qian & Schedl, 2004, r = .84; 64% in Stæhr, 2009, r = .80), the variance that could have been predicted by depth was already predicted by size. As a result, the predictive power of depth appears much lower than size. Therefore, the small proportion of variance explained by depth does not indicate that depth is less important for predicting reading and writing skills. To the contrary, when depth was the first variable entered into the regression equation, it could predict a much higher proportion of reading and listening variance, while size added only a small percentage (59% depth and 8% size in Qian, 2002; 55% and 6% in Qian & Schedl, 2004; 42% and 9% in Stæhr, 2009). These results suggest that size and depth in fact predict reading and listening proficiency in similar ways. Unlike reading and listening, however, the relative contributions of size and depth to speaking and writing skills remain unclear. In the present article, this constitutes the basis for conducting Study 1.

An additional concern is that the three abovementioned studies—Qian (2002), Qian and Schedl (2004), and Stæhr (2009)—all used the Word Associates Test (WAT) format (Read, 1993), which is designed to assess synonyms and collocations. According to Schmitt (2012), relationships between size and depth can vary depending on what specific areas of depth researchers target. The similar relationships of size and depth to reading and listening skills that these studies suggest may be due to their use of the same test format, and perhaps also because the synonyms that the WAT assessed overlapped with the size aspect. Therefore, studies employing formats different from the WAT are desirable.

Regarding the predictive power of lexical processing speed in relation to size, Van Gelderen et al. (2004) showed that size and speed were both moderately correlated with L2 reading comprehension (r = .63 and –.47, respectively), and that size was more effective (40%) than speed (22%) in predicting L2 reading, when each was separately entered into the regression equation. This pattern has also been observed in studies of L2 writing (Schoonen et al., 2003) and L2 speaking (De Jong et al., 2012, in press). Previous studies suggest that, unlike the case of size and depth, where their degree of predictive power is roughly the same, the predictive power of speed, albeit still substantial, is smaller. One reason for this may be the existence of a threshold level of speed: Speed may be strongly related to reading, writing, and speaking proficiency until learners reach a certain threshold level of sufficient speed, after which point further increase in speed does not entail greater speaking proficiency.

To conclude, size seems to hold considerable power in predicting L2 proficiency, when it is the first variable entered into the regression equation, while depth and speed contribute limited predictive powers for the remainder of the proficiency. However, when depth or speed is entered into the regression first, depth tends to exhibit a predictive power similar to size, whereas speed may have a predictive power less than size. This indicates the complicated nature of the contribution that these three lexical aspects make to language proficiency; thus far, however, only a limited number of studies have investigated this issue. To our knowledge, only Uenishi (2006) has tested the three aspects separately in relation to speaking, and even Uenishi's study is limited to novice speakers, and furthermore does not report test reliability or details about the tests and analysis. Thus, the report of Study 2 presented in this paper, inspecting the relationships between the factors of size, depth, speed, and L2 speaking proficiency, fills a significant gap in current research.

B. Relationships between L2 Vocabulary Knowledge and L2 Speaking

The well-known models of the speaking process proposed by Levelt (1989) and Kormos (2006) describe three main stages of speech production: conceptualization, formulation, and articulation. During the first stage, speakers form preverbal messages in the conceptualizer. In the formulator, they search for and retrieve necessary vocabulary from the mental lexicon, which contains information related to vocabulary and syntactic structures, in order to produce utterances with syntactic and phonological information. In the final stage, they utter the speech that they have formulated. Levelt stated that L1 speakers conduct these processes in parallel and automatically, without using substantial cognitive resources. However, L2 speakers experience much greater difficulty in executing such processes, a fact that prompted Kormos (2006) to propose an L2 speaking model.

According to both models, vocabulary holds a central position in formulating an utterance with the appropriate meanings, although other types of knowledge, including syntactic, morphological, and phonological knowledge, as well as nonlinguistic world knowledge and communication strategies, are also indispensible. The models indicate further the necessity of size, depth, and processing speed of vocabulary knowledge in speaking, because speakers use both form-meaning links (i.e., size) and the syntactic and morphological information associated with each word in the mental lexicon (depth), and because automatic, or at least relatively fast, lexical retrieval (speed) is required for smooth and effective communication.

In addition to the theoretical importance of vocabulary for speaking, empirical studies have been conducted into the relationship between vocabulary knowledge and speaking. Table 1 summarizes nine previous studies that have quantitatively investigated the relationships between L2 vocabulary knowledge and speaking. We did not include studies into the relationships between vocabulary knowledge and lexical complexity, because of the difficulty in identifying measures that can be interpreted with high validity when analyzing short texts (see Koizumi & In'nami, 2012).

TABLE 1.
PREVIOUS STUDIES ANALYZING RELATIONSHIPS BETWEEN VOCABULARY AND SPEAKING

Study	L1; L2; L2 level	Vocabulary aspect; test format	Speaking aspect; measure [speaking task used]	Statistical method and main results
Ishizuka (2000), *n* = 26	L1: Japanese; L2: English; novice[a]	(a) Depth; Word Associates Test format (Read, 1993)[b]	(1) Overall scores; composite scores of analytic rating scales [Eiken interview format]	Correlation (a & 1): *r* = .43
Segalowitz & Freed (2004), *n* = 40	L1: English; L2: Spanish; novice to advanced[a]	(a) L2 speed of lexical access; reaction time (RT) with L1 speed partialled out[c] (b) L2 efficiency of lexical access; (*SD* of RT)/(that person's mean RT) with L1 efficiency partialled out[c]	(1) Speed fluency; mean run length containing no filled pauses (e.g., *um*, *ah*) (2) Whether or not there was a gain in terms of the Oral Proficiency Interview scores between pretest and posttest, with an interval of 13 weeks [oral interview]	(a & 1) *r* = .38 (b & 1) *r* = .38 (a) or (b) explained by (2): both η^2 = .12
Koizumi (2005), *n* = 138	L1: Japanese; L2: English; novice	(a) Size; write L2 forms corresponding to L1 meanings (α = .91)	(1) Overall scores; composites of analytic rating scales (e.g., Task fulfillment; α = .86) [e.g., self-introduction, picture description]	Correlation (a & 1) *r* = .77
Uenishi (2006), *n* = 36	L1: Japanese; L2: English; novice[a]	(a) Size; Eiken vocabulary section[d] (b) Depth (word association); Lex30[e] (c) Access speed; reaction time to utter L2 forms corresponding to pictures	(1) Overall scores; composites of analytic rating scales (e.g., Content, Fluency, and Pronunciation; interrater reliability = .54) [describing a picture sequence] (2) Overall scores (interrater reliability = .73) [talking about hobbies]	Correlation (a & 1) *r* = .53; (a & 2) *r* = .48; (b & 1) *r* = .30; (b & 2) *r* = .12; (c & 1) *r* = –.27; (c & 2) *r* = –.01
Funato & Ito (2008), *n* = 62	L1: Japanese; L2: English; novice to intermediate[a]	(a) Size; write L2 forms corresponding to L1 meanings (b) Size; write L1 forms corresponding to L2 meanings	(1) Overall scores; composites of analytic rating scales (e.g., fluency, volume, and grammatical accuracy) [describing a comic and a picture]	Correlation (a & 1) *r* = .35 (b & 1) *r* = .27
Hilton (2008), *n* = 47	L1 French, German, and others; L2: English, Italian, and French; novice to advanced[a]	(a) Size; DIALANG (probably yes/no format; no details provided)	(1) Speed fluency; words per minute (2) Speed fluency; mean length of run (3) Repair fluency; mean length of hesitation (4) Repair fluency; percentage of production time spent hesitating (5) Repair fluency; rates of hesitation (6) Repair fluency; rates of retracing (7) Syntactic complexity; mean length of utterance (8) Accuracy; errors per 1,000 words[f]	Correlation (a & 1) *r* = .58; (a & 2) *r* = .67; (a & 3) *r* = –.39; (a & 4) *r* = –.55; (a & 5) *r* = –.66; (a & 6) *r* = –.52; (a & 7) *r* = .43; (a & 8) *r* = –.66
Milton et al. (2010), *n* = 30	L1: Arabic, Chinese, and others; L2: English; intermediate to relatively advanced	(a) Size (orthographic size); *X_Lex*; present L2 forms and check if test-takers think they know the meaning; yes/no format (b) Size (phonological size); *AuralLex*; yes/no format[g]	(1) IELTS speaking scores [oral interview]	Correlation: (a & 1): r_s = .35; (b & 1): r_s = .71 Linear regression: (1) explained by (b): R^2 = .42 Binary logistic regression: (1) explained by (b): Nagelkerke R^2 = .61
De Jong et al. (2012), *n* = 181	L1: 46 different languages (e.g., German); L2: Dutch; intermediate to advanced	(a) Size and depth (collocation) combined[h] (b) Speed of lexical retrieval; time to utter L2 forms in the picture-naming task (both α > .86)	(1) One latent speaking proficiency, rated on functional adequacy (α > .86) [descriptive and argumentative tasks]	SEM: (a & 1) *r* = .79; (b & 1) *r* = –.49; SEM multi-group analysis: Analyzed for each High and Low groups (*n* = 73, each); (1) explained by (a): High: R^2 = .45; Low: R^2 = .34; (1) explained by (b): High: R^2 = .04; Low: R^2 = .08
De Jong et al. (in press), *n* = 179	Same as De Jong et al. (2012)	Same as De Jong et al. (2012)	(1) Breakdown fluency; No. of silent pauses per 100 words (α = .96); (2) Breakdown fluency; mean silent pause duration (α = .93); (3) Breakdown fluency; No. of filled pauses per 100 words (α = .97); (4) Repair fluency; No. of corrections per 100 words (α = .77); (5) Repair fluency; No. of repetitions per 100 words (α = .91); (6) Speed fluency; mean duration of syllable (α = .97)[i]	Correlation (a & 1) *r* = –.39; (a & 2) *r* = –.02; (a & 3) *r* = –.33; (a & 4) *r* = –.43; (a & 5) *r* = –.24; (a & 6) *r* = –.58; (b & 1) *r* = .20; (b & 2) *r* = .16; (b & 3) *r* = .32; (b & 4) *r* = .25; (b & 5) *r* = .16; (b & 6) *r* = .32

Note. [a]Not reported, but evaluated by the authors according to the tests used. [b]Select L2 synonyms or collocates corresponding to L2 forms presented. [c]Semantic classification task, selecting either living or nonliving for L2 forms. [d]Select L2 forms appropriate to the sentential context. [e]Write L2 forms associated with L2 forms. [f][describing a video sequence]. [g]Listen to L2 forms and check if test-takers think they know the meaning. [h]Write L2 forms appropriate to the sentential context. [i][descriptive and argumentative tasks].

For instance, De Jong et al. (2012) investigated to what degree "L2 knowledge skills" and "L2 processing skills" explain L2 speaking proficiency (specifically functional adequacy), and whether the contributions of linguistics skills are different between more and less successful L2 learners. They administered eight speaking tasks and nine tests of linguistic skills to 181 adult learners of Dutch at intermediate and advanced levels, including a test of vocabulary knowledge (combining size and depth) and another of speed of lexical retrieval. They assessed size by requiring participants to supply L2 single-word forms appropriate to the sentence context, with one letter provided as a hint (90 items), and depth (specifically collocation) through a format that elicited L2 "prepositional phrases and verb-noun collocations" appropriate to the sentence (p. 17; 26 items). However, although they combined two formats to produce the total vocabulary knowledge scores, the depth items accounted for only 22% (26/116). Thus, we consider that their vocabulary knowledge test assessed mostly the aspect of size. In addition, a test of lexical retrieval speed measured the time it took participants to produce L2 forms corresponding to pictures provided. SEM analysis showed that vocabulary knowledge (size and depth combined) and intonation rating predicted 75% of speaking proficiency, and that speed contributed little to the prediction.

Reviewing these studies, we found varied results that may be explained by three main factors. First, the nine studies in Table 1 differed in the tasks/tests they administered and the aspects of vocabulary knowledge and speaking that they targeted. In terms of vocabulary aspects, four studies assessed size only (Funato & Ito, 2008; Hilton, 2008; Koizumi, 2005; Milton et al., 2010), with two studies measuring depth or speed (Ishizuka, 2000; Segalowitz & Freed, 2004) and three studies integrating size, depth, and processing speed (De Jong et al., 2012, in press; Uenishi, 2006). Regarding speaking aspects, six studies assessed overall speaking proficiency (De Jong et al., 2012; Funato & Ito, 2008; Ishizuka, 2000; Koizumi, 2005; Milton et al., 2010; Uenishi, 2006), while three assessed fluency (De Jong et al., in press; Hilton, 2008; Segalowitz & Freed, 2004). Strong correlations were found between size and overall speaking proficiency in three studies (e.g., r = .79 in De Jong, 2012) but not in two other studies (e.g., r = .27 in Funato & Ito, 2008). Additionally, weak or moderate correlations were found in some combinations: for example, between size and oral fluency in most studies (e.g., r = .67 in Hilton, 2008).

Second, some studies (e.g., Funato & Ito, 2008) failed to report their test and/or rater reliability, and a large measurement error may have led to underestimation of the strengths of relationships. SEM is a more appropriate tool than correlation or regression analysis for modeling relationships between variables with measurement error controlled for, in order to obtain rigorous and trustworthy results. Among the nine previous studies, only De Jong et al. (2012) used SEM, and they demonstrated strong relationships between size and overall speaking proficiency (r = .79).

Third, participants in the previous studies had different ranges of proficiency: novice only (e.g., Uenishi, 2006), novice to intermediate (Funato & Ito, 2008), novice to advanced (e.g., Hilton, 2008), and intermediate to advanced (e.g., De Jong et al., 2012). These differences in proficiency levels may have affected the results. For example, among five studies into associations between size and overall speaking proficiency, the two that used only intermediate and advanced learners showed high correlations (r = .79 in De Jong et al., 2012), whereas three that included learners at the novice level reported weak, moderate, or strong correlations (e.g., r = .53 in Uenishi, 2006). According to De Jong et al. (2012), a relatively wide range of proficiency levels should be incorporated when modeling proficiency, and studies dealing with only novice learners may not have sufficient variation, perhaps leading to weaker correlations. Further, the relative contribution model (e.g., Adams, 1980) posited that vocabulary plays a more important role in speaking proficiency among lower-level learners and that the impact of vocabulary becomes weaker as proficiency levels rise. This suggests that the contribution of vocabulary knowledge to speaking would be stronger among novice and intermediate than intermediate and advanced learners.

The mixed results generated by previous studies warrant further research into the relationships between L2 vocabulary knowledge and speaking proficiency, and particularly into the relative contribution of size, depth, and speed to L2 speaking proficiency. This article attempts to cover wider aspects of vocabulary knowledge (size, depth, and speed) and speaking (overall speaking, fluency, accuracy, and syntactic complexity [SC]), using SEM to account for measurement error, and including learners of a relatively wide range of proficiency levels. We employ novice- and intermediate-level learners, and compare our results with those of De Jong et al. (2012), who employed intermediate- and advanced-level learners, in order to examine the relative contribution model (e.g., Adams, 1980).

C. Present Study

Although vocabulary knowledge is only one among the many variables that affect oral production (De Jong et al., 2012, in press), the literature review above shows that it is theoretically indispensable. However, although empirical investigations have generally supported this, the limited number of studies conducted justifies further research. This study conceptualizes vocabulary knowledge according to three aspects: size, depth, and speed. We also followed Housen and Kuiken (2009) in regarding speaking proficiency as consisting primarily of fluency, accuracy, and SC.

We conducted two studies: Study 1 examines the relationships between size, depth, and speaking proficiency, while Study 2 adds speed to the design of Study 1. Our research question is to what extent L2 speaking proficiency is predicted by L2 vocabulary knowledge, in terms of overall knowledge, size, depth, and speed. Drawing on previous studies (e.g., De Jong et al., 2012; Milton et al., 2010; Qian & Schedl, 2004), we hypothesize that vocabulary knowledge contributes substantially to the prediction of speaking proficiency, and that size and depth predict speaking similarly to each other, and to a greater degree than speed.

II. Study 1

A. Method

1. Participants

The participants were 224 Japanese native speakers, who had studied English as a foreign language for two to five years. They were secondary school learners (from 14 to 18 years old), including 97 males and 127 females. They were judged to be proficient at novice and lower-intermediate levels ("below A1" to B1 levels), based on their self-reported grade in the Eiken Test, which was translated by means of the conversion table (STEP, 2012) into the levels of the Common European Framework of Reference for Languages. Out of a larger pool of data, we selected those who took vocabulary tests and a speaking test and uttered at least one clause for every speaking task.

2. Instruments

Vocabulary knowledge was elicited in a decontextualized, controlled manner using a paper-and-pencil format. Vocabulary tests covered four aspects: size, derivation, antonym, and collocation, with the latter three sections assessing depth (see Table 2 for test formats). The three aspects of depth were selected as follows: (a) since knowledge of word association is essential in activating words, connecting them, and forming an utterance, common and typical word association responses—antonym and collocation (Aitchison, 2003)—were selected; (b) derivation was chosen to encompass the wider aspects of vocabulary knowledge.

TABLE 2.
EXAMPLES OF VOCABULARY TEST ITEMS IN STUDY 1

Size Test (78 items)
Write the English word that best corresponds to the Japanese meaning on your answer sheet.
2. ネズミ [*nezumi*] (m) [Answer: *mouse* (*mice*)]
Derivation Test (20 items)
Change the form of each English word below according to the part of speech provided in [ ]. Write only one word. Do not write words with *-ing* or *-ed*.
10. supporter [Verb: do the action of …] ( ) [*support(s)*]
Antonym Test (17 items)
Write one word that has the opposite meaning to the word presented.
6. start ( ) [Example answers: *end, finish, stop, termination*]
Collocation Test (18 items)
Write one English word that fits ( ) (a noun).
4. wash (a/an/the) ( ) [e.g., *dish(es), hand, mouth, car*]

Note. The instructions were written in Japanese and included examples.

In the size test, 78 words were randomly selected from the 3,000 most frequent lemmas in the JACET8000, a word list specifically tailored for Japanese learners of English (JACET Basic Word Revision Committee, 2003). In the derivation test, 20 derivational suffixes were selected. In the antonym and collocation tests, the words selected had at least one possible answer belonging to the 3,000 most frequent lemmas in the JACET8000.

The 15-minute, tape-mediated speaking test required test-takers to produce real-time monologues. They were not given pre-task planning time. The test included five tasks: a self-introduction task (Task 1), two tasks describing a single picture (Tasks 3 and 4), and two tasks explaining the differences between two pictures (Tasks 2 and 5).

3. Procedures and Analyses

Test-takers took four vocabulary tests in the following order: size, derivation, antonym, and collocation. The speaking test was conducted a week before or after administering the vocabulary tests.

The vocabulary tests were dichotomously scored. Scoring criteria were developed for the depth tests using seven dictionaries. In the antonym and collocation tests, responses that did not match the criteria—unless completely incorrect—were judged by three raters. The internal consistency of the three raters was moderate and considered acceptable (α = .65 for antonym; α = .63 for collocation). Responses on which the three raters disagreed were evaluated by two additional raters and were scored as correct when three of the five raters agreed. The five raters comprised three native English speakers and two Japanese advanced English learners. The reliability estimates of the four vocabulary tests were high (α = .73 to .92).

Utterances produced in the speaking test were transcribed for 45 seconds for each task, and then coded in terms of features such as the number of AS-units (Analysis of Speech units; Foster, Tonkyn, & Wigglesworth, 2000). With regard to the number of error-free clauses, one third of the utterances were evaluated by four raters (i.e., two native English speakers and two advanced English learners). The reliability was high overall (α = .88 to .93). After clarifying the judgment rule for errors, the remainder of the transcripts were judged by two raters (α = .92 to .98). Points of disagreement were discussed until a consensus was reached.

Study 1 conceptualized speaking proficiency as consisting of fluency, accuracy, and SC, each element of which was represented by aspects of performance that were measured through the five tasks. Fluency can be classified into three fluency dimensions (Tavakoli & Skehan, 2005): speed fluency (measured by, for example, speech rate), repair fluency (assessed according to indices of self-correction, repetition, false starts, and replacements), and breakdown fluency (evaluated by pause-related measures). We focused on speed and repair fluency, because of the poor conditions of our recordings, and on the assumption that our decision would not greatly affect our results, as previous studies have shown

that speed fluency is often closely associated with breakdown fluency, and that the two load on the same factor in factor analyses (e.g., Tavakoli & Skehan, 2005).

We used four discourse analytic measures: the number of tokens per minute for speed fluency (where "tokens" refers to pruned tokens after the exclusion of dysfluency markers); the number of dysfluency markers (i.e., functionless repetitions, self-repairs, and filled pauses, such as *mm*, *ah*) per minute for repair fluency; the number of error-free clauses per clause for accuracy; and the number of clauses per AS-unit for SC. We selected these four measures because they have been used often in previous research (e.g., Segalowitz & Freed, 2004; Tavakoli & Skehan, 2005), and because they were highly correlated with other, similar measures.

To perform SEM analyses, we employed syntax for EQS (Version 6.1; Bentler, 2010) and Amos (Version 7.0.0; Arbuckle, 2006) for visual display. Based on Byrne (2006), we checked univariate and multivariate normality by examining whether the *z* scores of skewness and kurtosis values were within |3.30| ($p < .01$; Tabachnick & Fidell, 2007), and whether Mardia's normalized estimate values were 5.00 or less (Byrne, 2006). To estimate model parameters, the robust maximum likelihood method was employed, because some of the variables were nonnormally distributed. One factor loading from each factor was fixed to 1.00 for scale identification. The following model fit indices were used: the comparative fit index (CFI) of 0.90 or above (Arbuckle & Wothke, 1995), the root mean square error of approximation (RMSEA) of 0.08 or below, and the standardized root mean square residual (SRMR) of .08 or below (Hu & Bentler, 1999). We had no missing data, and obtained a sufficient sample size for SEM, that is, over 200 (Kline, 2010). Descriptive statistics of scores and measure values shows that there were some variations in vocabulary and speaking measures. Results suggest that all the variables were correlated to some degree ($r = –.13$ to .80) and that there were no very high correlations (more than $r = |.90|$), which cause problems of multicollinearity (Tabachnick & Fidell, 2007).

B. Results

To examine the structures of each factor, two models of vocabulary knowledge were tested (see Figure 1). In Model 1, the vocabulary knowledge factor subsumed size, derivation, antonym, and collocation. Model 2 had one depth factor, which was correlated with the observed variable of size. Both models were equally acceptable (e.g., CFI = 1.00; RMSEA = 0.07 [90% confidence interval: 0.00, 0.07]; SRMR = .01) and statistically indistinguishable. The correlation in Model 2 between the depth factor and the size variable was very high ($r = .94$), suggesting that size and depth can be considered one construct. Thus, we selected Model 1. This strong association between size and depth is in line with previous research (e.g., Akbariana, 2010). The high factor loadings from the factor to each observed variable (β = .72 to .94) indicated that the four vocabulary test variables effectively assessed vocabulary knowledge.

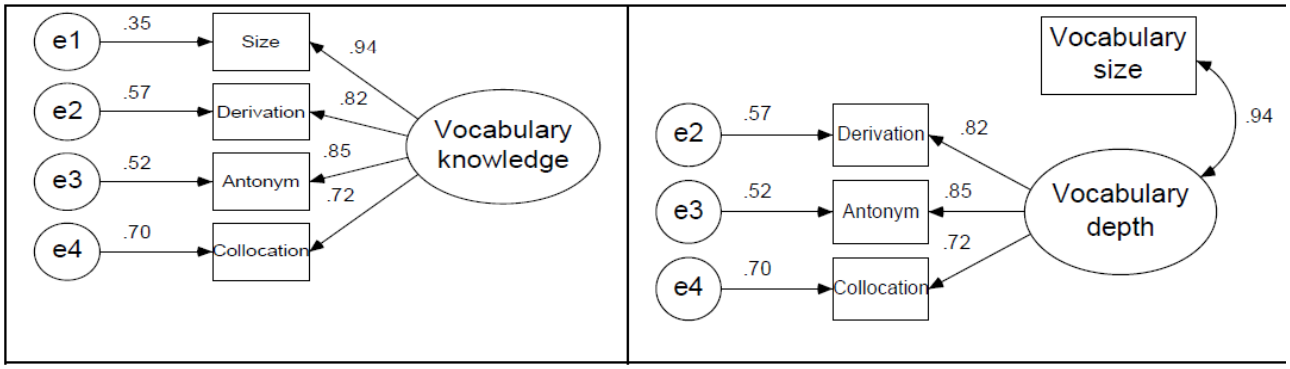

Figure 1. Model 1 (left): One-factor model of vocabulary knowledge. Model 2 (right): Size and depth model.

Regarding speaking proficiency, we tested a four-factor correlated model (Model 3), in which five observed task variables underlay four correlated factors representing the four aspects of speaking proficiency (i.e., speed fluency, repair fluency, accuracy, and SC). Model 3 fit the data well (e.g., CFI = .93; RMSEA = 0.05 [0.04, 0.06]; SRMR = .06). Other competing models did not fit the data well, such as one with a higher-order speaking proficiency factor represented by the four factors (e.g., CFI = .84; SRMR = .17) and another with a unitary speaking proficiency factor without any CAF factors (e.g., CFI = .73; SRMR = .10). In Model 3, there were substantial correlations between the four factors ($r = .35$ to .88), except for the relationship between repair fluency and accuracy ($r = .13$). Furthermore, we observed generally substantial factor loadings from each factor to each task variable (β = .40 to .88), with the exception of Task 2 Accuracy, Task 1 SC, and Task 3 SC (β = .20 to .30). This suggested that the factors of speed fluency, repair fluency, accuracy, and SC were, in general, effectively measured by the variables used.

Once the models of vocabulary knowledge (Model 1) and speaking proficiency (Model 3) had been found to fit the data, Model 4 (see Figure 2) was created to test the research question, in which (a) the vocabulary knowledge factor is hypothesized to predict the speed fluency, repair fluency, accuracy, and SC factors, and (b) there are hypothesized correlations between the measurement errors of four latent factors of speaking proficiency (represented by "D" in Figure 2), which can be interpreted as variances not explained by vocabulary knowledge. Hypothesizing correlations between the measurement errors was sensible because weak or moderate relationships between the four latent factors of

speaking proficiency (as found in Model 3) suggest the divisibility of the speaking components, with some of their variances unexplained by vocabulary knowledge. Model 4 showed the good fit (e.g., CFI = .95; RMSEA = 0.04 [0.03, 0.05]; SRMR = .06). An alternative model that is the same as Model 4 but without correlations between the errors did not fit the data well (e.g., SRMR = .10).

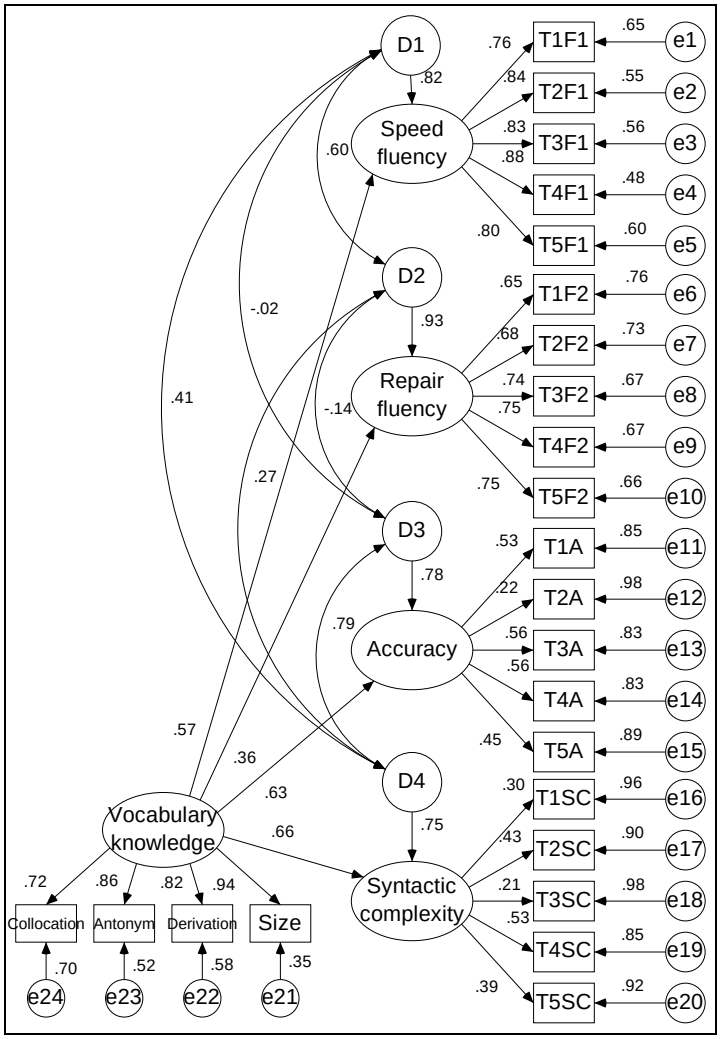

Figure 2. Model 4: Effects of vocabulary knowledge on speaking proficiency. All the testable path and correlation coefficients were significant, except for the correlation between "D1" and "D3" and between "D2" and "D3." F1 = Speed fluency. F2 = Repair fluency. A = Accuracy.

According to Model 4, vocabulary knowledge predicted aspects of speaking proficiency. It predicted speed fluency strongly (β = .57), explaining 32% of the speed fluency factor variance; repair fluency moderately (β = .36), with 13% explained; accuracy strongly (β = .63), with 40% explained; and SC strongly (β = .66), with 44% explained. Correlations between the errors varied from strong (i.e., *r* = .79, between errors of accuracy and SC), moderate (e.g., *r* = .60, between errors of speed fluency and repair fluency), to almost zero (e.g., *r* = –.02, between errors of speed fluency and accuracy). It should be remembered that these values suggest the strengths of relationships when other variables are held constant.

As for the relative contribution of size and depth, the following formula can be used to show the degree to which two variables are related: multiply the loadings of paths between the two variables. For example, the loading of the speed fluency factor, as predicted by the size variable, was β = .54, obtained by multiplying the loading from the vocabulary factor to the size variable (β = .94) by the loading from the vocabulary factor to the speed fluency factor (β = .57). The loading squared (.54*.54) indicates the proportion of the variance explained (29%; see Table 3). This means that the speed fluency factor was largely explained by size (29% out of 32%), with a small proportion (the remaining 3%) explained by other vocabulary variables. This trend applied for repair fluency (11% vs. 2%), accuracy (35% vs. 5%), and SC (38% vs. 6%) factors. Moreover, depth (derivation, antonym, and collocation) predicted similar proportions of variance of speaking proficiency, when it was considered first in the prediction (speed fluency: 17% to 24%; repair fluency: 7% to 10%; accuracy: 21% to 29%; SC: 23% to 32%).

The models used in Study 1 included size and depth variables. Study 2, meanwhile, added speed, in order to further examine the relationship between vocabulary knowledge and speaking proficiency.

TABLE 3.
PROPORTIONS OF SPEAKING PROFICIENCY EXPLAINED BY VOCABULARY KNOWLEDGE IN MODEL 4

Predicted variable	Predicted by	Vocabulary knowledge		Size		Deri-vation		Anto-nym		Collo-cation	
		β	R^2	β	R^2	β	R^2	β	R^2	β	R^2
Speed fluency		.57	.32	.54 (.94*.57)	.29 ((.54)^2)	.47	.22	.49	.24	.41	.17
Repair fluency		.36	.13	.34 (.94*.36)	.11 ((.34)^2)	.30	.09	.31	.10	.26	.07
Accuracy		.63	.40	.59 (.94*.63)	.35 ((.59)^2)	.52	.27	.54	.29	.45	.21
SC		.66	.44	.62 (.94*.66)	.38 ((.62)^2)	.54	.29	.57	.32	.48	.23

III. STUDY 2

A. Method

1. Participants

Study 2 analyzed 87 test-takers, who took three vocabulary tests and a speaking test, and who uttered at least one clause during an opinion-statement task (see below). All the participants were Japanese native speakers studying English, with 49 undergraduate (56%) and 38 graduate students, from 16 Japanese and 1 British university. There were 57 females (66%) and 24 males, with 72 English-majors (83%) and 15 non-English-majors (e.g., international relations). According to Pearson Education, Inc. (2008), 98% of the test-takers (n = 86) belonged to either novice or intermediate levels ("below A1" to B2), and one belonged to an advanced level (C1).

2. Instruments

Three computer-based vocabulary tests (Mochizuki et al., 2010; see Figure 3) and a telephone-based speaking test were used. One vocabulary test (JACET8000 Vocabulary Size Test; J8VST) aimed to assess size up to 5,000 lemma on the basis of the JACET8000. For each 1,000 lemma level, 25 words were randomly selected. This 125-item multiple-choice test required test-takers to select the L2 form that corresponded most closely to the L1 meaning provided. Each question included five options, one of which was labeled "I don't know," to reduce random guessing behavior.

A second vocabulary test, a Lexical Organisation Test (LOT; Flash Version), was designed to test lexical organization, or more specifically the strength of collocation, which was considered an aspect of depth in this study. Test-takers were asked to select the strongest collocation from three choices (e.g., *dark mouth*, *dark horse*, and *horse mouth*). Although all three combinations are possible, *dark horse* is the correct answer, since the two words are most strongly connected. The LOT contained 50 items. All the target word combinations were ones that native English speakers could usually identify by intuition but that L2 learners would have to learn. Each test item consisted of two collocations (identified by the Cobuild Collocation Sampler) and one two-word, non-collocation string. One of the two collocations became the answer on the basis of results indicating that at least 85% of 20 native English speakers agreed on a stronger connection between the two words. At least one word in the correct collocation belonged to the 1,000 most frequent lemmas in the JACET8000.

A third vocabulary test, a Lexical Access Time Test (LEXATT), was intended to assess how quickly test-takers could recognize word form and meaning. It consisted of two tasks for each of 40 items. First, reaction time was assessed by measuring the time from when a target word was presented until test-takers indicated that they recognized the form and meaning by releasing a pushed button. Second, test-takers selected one L1 meaning out of two options, corresponding to the L2 form they had seen. The second task aimed to ensure that they had actually understood the meaning of the L2 form. All the words presented had four letters and were selected from the 3,000 most frequent lemmas in the JACET8000. For the LOT and LEXATT, three and five practice items were presented, respectively, before the test.

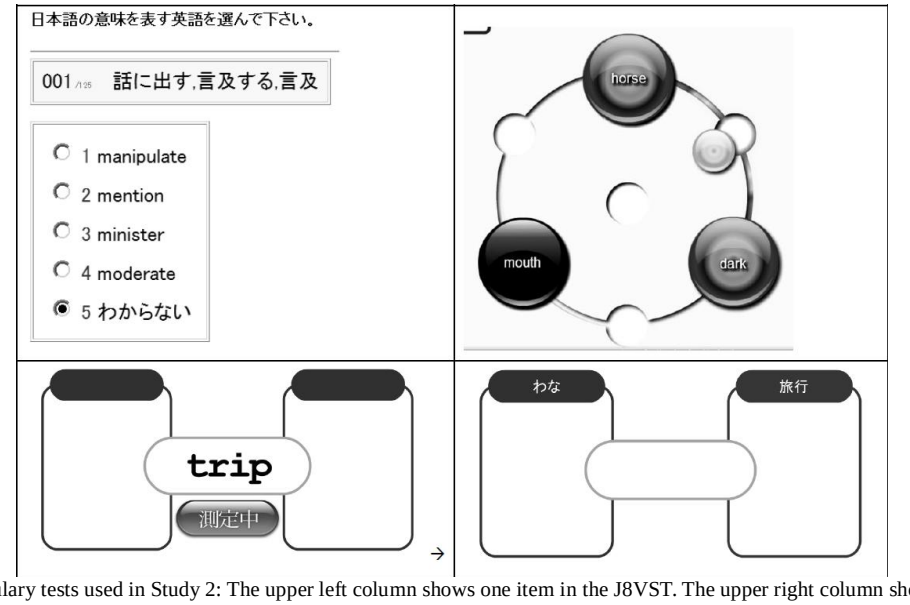

Figure 3. Vocabulary tests used in Study 2: The upper left column shows one item in the J8VST. The upper right column shows the LOT. The bottom left and right columns show the two tasks of the LEXATT.

In order to test speaking proficiency, the Versant™ English Test (Versant, hereafter; Pearson Education, Inc., 2008) was administered. This 15-minute, technology-mediated test included five tasks: (a) reading sentences aloud, (b) repeating sentences, (c) answering commonsensical questions with a few words, (d) reordering three blocks of phrases into an understandable sentence (e.g., test-takers hear "*was reading*," "*my mother*," and "*her favorite magazine*," and should respond, "*My mother was reading her favorite magazine.*"), and (e) stating opinions. For each task, test-takers both listened to questions and responded in English. Their responses were recorded, and scores were produced through an automated scoring system, on the basis of responses to the first four tasks. The Versant is designed to elicit "facility in spoken English" (Pearson Education, Inc., 2008, p. 7) and efficiency of processing spoken language, which is an essential aspect of speaking proficiency (Van Moere, 2012).

In Study 2, we used (a) the Versant Overall Score (ranging from 20 to 80), which was generated based on the first four tasks, and (b) the discourse analytic measures of speed fluency, repair fluency, accuracy, and SC, which were derived from utterances in the final task (opinion-statement). This task had an open-ended format, and required test-takers to listen to a prompt twice, before expressing their opinions regarding family life or personal choices for 40 seconds, without planning time. An example of the questions asked is, *Do you think television has had a positive or negative effect on family life? Please explain.* (Pearson Education, Inc., 2008, p. 6). While this task consisted of three prompts, we chose to analyze the second prompt only, for two reasons: First, some test-takers did not fully understand the format of the first prompt, and were unable to speak to the full extent; second, some test-takers, especially those with higher proficiency, tended to speak less in the third than in the second prompt, probably because they had grasped by this point how briefly they were required to speak. Hence, the second prompt elicited, on average, the longest and most varied speech.

For the second prompt, each test-taker responded to one of 48 questions presented in the Versant. Although the utterances did not seem to differ greatly, because of the similar nature of the topics, the fact that not all test-takers answered the same questions (e.g., student 1 answered question 1; student 2 answered question 3) may limit the generalizability of findings from Study 2.

3. Procedures and Analyses

Test-takers took three vocabulary tests in the following order: LEXATT, J8VST, and LOT. They also took the Versant within one month of the vocabulary tests. Before they took the Versant, they were instructed to read the instructions carefully and to practice the task formats by using the example questions.

The J8VST and LOT were scored dichotomously. For the LEXATT, the response time was analyzed only when test-takers selected the right answer in the second (meaning confirmation) task. The response time for each item was averaged; lower values for the LEXATT indicate a faster speed for recognition of word form and meaning. The internal consistency was high for both the J8VST (α = .96) and the LOT (α = .77), but could not be calculated for the LEXATT, on account of the lack of item level data.

After we had transcribed utterances for 40 seconds in the open-ended Versant task, we counted the number of occurrences of certain features (e.g., clauses). Two raters (Japanese advanced English learners) judged one-third of the utterances in terms of error-free clauses; interrater reliability was high (agreement ratio = .87; κ = .69, $p < .001$). We resolved rater disagreements through discussion, and then one rater judged the remainder of the transcripts. We computed four discourse analytic measures—the same as those used in Study 1—to assess speed fluency, repair fluency, accuracy, and SC.

In Study 2, we regard speaking proficiency as consisting of processing efficiency, speed fluency, repair fluency, accuracy, and SC, each of which was indicated by the Versant Overall Score and the four discourse analytic measures. While the Versant Overall Score is derived by using decontextualized tasks to elicit short, sentence-level utterances, the four discourse analytic measures were computed based on a contextualized task requiring topic-based, discourse-level performance. By using the two types of speaking tasks, we intended to capture wider areas of speaking proficiency. It is important to note that although the scores and utterances are based on integrated speaking tasks, and may reflect some degree of listening ability, we interpreted them purely in terms of speaking proficiency.

SEM was conducted in the same manner as in Study 1. Since univariate normality was and found to be violated, the robust maximum likelihood method was employed for estimation. The data included no missing values. The sample size was 87, which was smaller than the minimum sample size of 100 (Kline, 2010), but could be acceptable as judged by the power analysis we conducted. Power analysis of SEM models can provide evidence for whether the sample size is sufficient for a model to have adequate precision of parameter estimates and statistical power. This study applied Monte Carlo power analysis (Muthén & Muthén, 2002), using Mplus (Version 6.12; Muthén & Muthén, 2011). The results suggested that parameter bias, standard error bias, coverage, and power were all adequate. For example, Model 7 had a power of .94 to 1.00, which satisfied the criterion of .80 or above. Therefore, the sample size in Study 2 (n = 87) was considered to be satisfactory. We found moderate correlations between the variables (r = –.51 to .80), which were not high enough to cause problems of multicollinearity.

B. Results

We first constructed two models: One model expressed a vocabulary knowledge factor with three vocabulary variables (size, depth, and speed; Model 5), and the other a speaking proficiency factor, reflected by five speaking variables (processing efficiency, speed fluency, repair fluency, accuracy, and SC; Model 6). We were unable to test the appropriateness of Model 5, because it was just statistically identified (i.e., its degree of freedom was zero and we could not compute fit indices). Model 6 fit the data well (e.g., CFI = .99; RMSEA = 0.07 [0.00, 0.18]; SRMR = .05). We then constructed a model in which the vocabulary knowledge factor was hypothesized to predict the speaking proficiency factor (Model 7; see Figure 4). We adopted this model because of its good fit with the data (e.g., CFI = .97; RMSEA = 0.08 [0.00, 0.13]; SRMR = .06). The standardized regression coefficients (β) from the vocabulary factor to each vocabulary variable were substantial, ranging from –.58 to .87. The standardized regression coefficients (β) from the speaking proficiency factor to each speaking variable were also high, ranging from .43 to .91. This means that both vocabulary and speaking factors were measured well by the observed variables. One unexpected result was that the repair fluency variable loaded positively on the speaking proficiency factor (β = .43). Since this measure was computed by the number of *dysfluency markers* per minute, higher values indicated lower repair fluency. This result suggests that learners with higher speaking proficiency tend not only to speak faster, with more accurate and more syntactically complex sentences, but also to produce more filled pauses (e.g., *um*, *ah*) in the utterance, with *more* repetitions and self-repairs. This could be explained by the participants' limited proficiency, in that they were probably unable to avoid repairing their utterances while searching for and uttering words at rapid speed (Wood, 2010).

Table 4 shows that in Model 7, vocabulary knowledge predicted 84% of speaking proficiency, 70% of efficiency of processing spoken language, 64% of speed fluency, 16% of repair fluency, 20% of accuracy, and 21% of SC. Moreover, size was found to predict 63% of speaking proficiency when it was entered first into the regression equation, with 21% explained by depth and speed [84% – 63%]), 52% of processing efficiency, 48% of speed fluency, 12% of repair fluency, 15% of accuracy, and 15% of SC. Furthermore, depth predicted speaking similarly to size: 60% of speaking proficiency, 50% of processing efficiency, 46% of speed fluency, 11% of repair fluency, 14% of accuracy, and 15% of SC. In contrast, speed predicted speaking less than both size and depth: 28% of speaking proficiency, 23% of processing efficiency, 21% of speed fluency, 5% of repair fluency, 7% of accuracy, and 7% of SC.

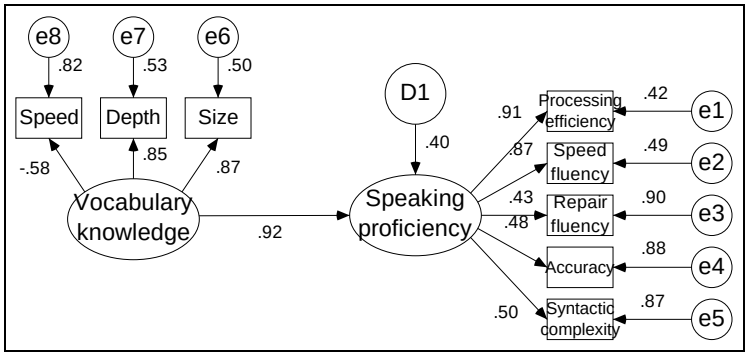

Figure 4. Model 7: Predicting speaking proficiency from vocabulary knowledge. All the testable path and coefficients were significant.

TABLE 4.
PROPORTIONS OF SPEAKING PROFICIENCY EXPLAINED BY VOCABULARY KNOWLEDGE IN MODEL 7

Predictor variable	Predicted by	VK		Size		Depth		Speed	
		β	R^2	β	R^2	β	R^2	β	R^2
Speaking proficiency		.92	.84 ((.92)^2)	.79 (.87*.92)	.63	.78	.60	–.53	.28
Processing efficiency		.83 (.92*.91)	.70 ((.83)^2)	.72 (.87*.92*.91)	.52	.71	.50	–.48	.23
Speed fluency		.80 (.92*.87)	.64 ((.80)^2)	.69 (.87*.92*.87)	.48	.68	.46	–.46	.21
Repair fluency		.40 (.92*.43)	.16 ((.40)^2)	.34 (.87*.92*.43)	.12	.34	.11	–.23	.05
Accuracy		.44 (.92*.48)	.20 ((.44)^2)	.38 (.87*.92*.48)	.15	.37	.14	–.25	.07
SC		.45 (.92*.50)	.21 ((.45)^2)	.39 (.87*.92*.50)	.15	.38	.15	–.26	.07

IV. OVERALL DISCUSSION

In Studies 1 and 2, L2 vocabulary knowledge was generally found to substantially explain L2 speaking proficiency and its various aspects, with a constant exception being repair fluency. These results suggest that learners at novice and intermediate levels with greater vocabulary knowledge in terms of size, depth, and speed, are likely to have higher speaking proficiency, enabling them to produce more rapid, accurate, and syntactically complex oral performance. Furthermore, when other variables were held constant, size and depth could predict considerable amounts of the variances in speaking proficiency that were explained by vocabulary knowledge and speed could predict speaking proficiency less than size and depth.

The finding that L2 vocabulary knowledge considerably predicts L2 speaking proficiency generally accords with most previous studies of L2 speaking, especially the one previous study that has used SEM (De Jong et al., 2012). Moreover, the indication that size and depth almost equally predict L2 speaking proficiency, produced in both Study 1 and Study 2, corroborates previous research into L2 reading and listening (e.g., Stæhr, 2009). If we recall that both our studies used formats different from the WAT for assessing depth, it becomes clear that these consistent findings suggest greater generalizability regarding similarities of size and depth in predicting L2 proficiency.

The finding in Study 2, that speed explained L2 speaking less than size, was also in line with previous research into L2 speaking, reading, and writing (De Jong et al., 2012; Schoonen et al., 2003; Van Gelderen et al., 2004). However, no study, including the current one, has yet distinguished nonlanguage, general motor speed from L2 speed. Thus, measures of L2 lexical processing speed that partial out nonlanguage speed (see Segalowitz & Freed, 2004) would further clarify the relationship between L2 speed and speaking proficiency.

Furthermore, a large proportion of speaking proficiency being explained by size alone was consistent with previous studies on L2 speaking, reading, listening, and writing (e.g., Hilton, 2008). This indicates that size could be a powerful single predictor of L2 proficiency. The proportion of speaking proficiency explained by size in Study 2 (63%) was notably similar to the one in De Jong et al. (2012; 62%). Since both studies used SEM, with one speaking proficiency factor posed in the model and fairly consistent results obtained, this may suggest that the proportion of speaking proficiency predicted by size could be similar across both novice to intermediate learners, as in our study, and intermediate to advanced learners, as in De Jong et al. This seems to contradict Adams's (1980) relative contribution model, which predicts that vocabulary has a greater effect on speaking among lower-proficiency learners. In fact, De Jong et al. (2012) showed that the explanatory power of vocabulary knowledge was stronger among advanced rather than intermediate learners (43% vs. 34%, respectively) and that the same patterns were found for grammatical knowledge, speed of sentence building, and pronunciation. Our findings as well as those of De Jong et al. (2012) suggest the need to revise the relative contribution model on the basis of empirical speech production data.

Although the results of Studies 1 and 2 showed consistently that speaking proficiency was substantially predicted by vocabulary knowledge, size, and depth, some differences were observed. For example, speed fluency was predicted by vocabulary knowledge less in Study 1 (32%) than in Study 2 (64%), whereas the opposite trend was observed in terms of accuracy and SC (e.g., 40% vs. 20% regarding accuracy). However, these differences are difficult to explain, because the two studies differed in several aspects: the proficiency ranges of target learners (i.e., wider ranges in Study 2), the aspects of vocabulary knowledge assessed (i.e., size, derivation, antonym, and collocation in Study 1, vs. size, collocation, and speed in Study 2), and speaking tasks (e.g., simple speaking tasks related to familiar topics in Study 1 vs. more complex, cognitively demanding tasks that required test-takers to listen and respond to prompts in Study 2). Further studies with these variables controlled for would clarify the different predictive powers of vocabulary knowledge toward fluency, accuracy, and SC. Nevertheless, the consistency of the results of Studies 1 and 2, and of previous studies also, provides strong evidence for considerable relationships between vocabulary knowledge and speaking proficiency.

The results from Study 1, Study 2, and previous studies (e.g., De Jong et al., 2012) suggest that vocabulary knowledge considerably contributes to speaking proficiency, including speed fluency, accuracy, and SC. This finding can be explained according to two different perspectives: proficiency and processing. First, learners who have higher overall L2 proficiency tend also to exhibit higher vocabulary knowledge and higher speaking proficiency. This allows them to produce faster, more accurate, and syntactically complex utterances. The second perspective can be explicated in terms of speaking models (Kormos, 2006; Levelt, 1989) and Skehan's (2009) theory. Skehan related fluency,

accuracy, and complexity of learner production to Levelt's speaking model, and argued that preverbal messages created at the conceptualizer stage affect complexity, and that operations at the formulator stage affect fluency and accuracy. He further explained that differences in speaking processes between native and nonnative speakers are partially attributed to the quantity and quality of vocabulary knowledge. The current study attempts to extend these frameworks, by explaining the relationships between vocabulary knowledge, fluency, accuracy, and SC.

Vocabulary knowledge and fluency may be associated with each other because L2 learners with larger and deeper vocabulary knowledge, and faster access to it, can perform lexical searches more easily and quickly. Learners with greater vocabulary knowledge can recall adequate words and use them for speaking through knowledge of antonyms and collocations (Aitchison, 2003). Consequently, processing will be smoother for them than for those with a smaller lexicon, although the speed of intermediate-level learners with greater vocabulary knowledge is still much slower than for high-proficiency learners, and their processing is far from automatic. On the other hand, learners with poorer vocabulary knowledge may not be able to find appropriate words, or may take longer to search for words at the formulation stage, resulting in reduced speed fluency.

The impact of vocabulary knowledge on accuracy and SC may be related to the ease of lexical searches. Since learners with larger and deeper vocabulary and faster lexical access can find words more easily, their cognitive resources, which are limited in terms of attentional capacity, remain available to attend to other areas, including processes in the conceptualizer and formulator. Cognitive processing space directed to the formulator enables speakers to produce more accurate utterances, while attention directed to the conceptualizer enhances SC. The latter scenario may require more explanation. Having some attentional resources available for areas other than lexical searches could possibly enable speakers to attempt to "formulate more complex ideas" (Skehan, 2009, p. 520), perhaps even prompting them to "repair for good language," or to monitor their language for "a more sophisticated manner of expression" (Kormos, 2006, p. 125). After comparing their capability to encode the language with the preverbal plan that they would like to express, they may create more complex preverbal messages, which are encoded in the formulator in more complex language. In contrast, those with smaller and less organized lexicons, and with slower access, tend to be occupied with retrieving appropriate words. Consequently, they may not be able to direct attention to speaking processes other than lexical retrieval (i.e., conceptualizing, formulating, articulating, and monitoring). This results in the production of less fluent, less accurate, and syntactically less complex utterances, and eventually of poorer speaking performance. This process highlights the importance of size, depth, and speed in speaking.

V. Conclusions

The research question concerned the extent to which L2 speaking proficiency is predicted by L2 vocabulary knowledge, size, depth, and speed, with a focus on Japanese learners of English at novice to intermediate levels. Across Studies 1 and 2, we found that speaking proficiency can be explained by vocabulary knowledge to a substantial degree (32% to 44% in Study 1, and 84% in Study 2), with the exception of for repair fluency (13% and 16%, respectively), and that a considerable degree of speaking proficiency can be explained by size alone (29% to 38%, and 63%, respectively) or by depth alone (17% to 32%, and 60%, respectively), and to a lesser degree by speed alone (28%). These results substantiate the importance of size, depth, and speed in speaking proficiency, and the effectiveness of size and depth, and of speed to a lesser degree, as predictors of speaking proficiency. Although this research showed that speaking proficiency could be effectively predicted by vocabulary knowledge, further experimental studies are necessary to examine whether enhancing vocabulary knowledge actually leads to an increase in speaking proficiency.

While this research presented evidence that vocabulary knowledge explains speaking proficiency, the results may be restricted to the design of the studies in this article, including the target learners, tests, and measures selected. Future research should include more aspects (e.g., ability to interact with interlocutors appropriately) and more measures (e.g., length of noun phrases and frequency of discourse markers, for SC measures).

The current study had four key strengths. First, rather than including a wide range of linguistic components hypothesized to affect speaking proficiency (like De Jong et al., 2012, in press), we focused on aspects of vocabulary knowledge as predictors, and separately assessed size, depth, and speed. This enabled us to investigate constructs of vocabulary knowledge in a more detailed manner. Second, we administered multiple tasks measuring distinctive aspects of speaking proficiency, reflected by fluency, accuracy, SC, and processing efficiency. Third, we modeled variables using SEM, while controlling for measurement error. Fourth, we targeted learners with a relatively wide range of proficiency, following the advice of De Jong et al. (2012), regarding the importance of including heterogeneous learner samples when constructing proficiency models.

Acknowledgment

This work was supported by Grand-in-Aid for Scientific Research (KAKENHI) of the Ministry of Education, Culture, Sports, Science and Technology in Japan (No. 22720216 and 19320084), and Knowledge Technologies, Pearson. We are deeply indebted to Akihiko Mochizuki, Masamichi Mochizuki, Toshihiko Uemura, Kazumi Aizawa, Naoki Sugimori, Shin'ichiro Ishikawa, Tatsuo Iso, and Shin'ichi Shimizu for their invaluable comments on earlier versions of this paper.

REFERENCES

[1] Adams, M. L. (1980). Five coocurring factors in speaking proficiency. In J. R. Firth (Ed.), *Measuring spoken language proficiency* (pp. 1–6). Washington, DC: Georgetown University Press.
[2] Aitchison, J. (2003). Words in the mind (3rd ed.). Malden, MA: Blackwell.
[3] Akbariana, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System, 38,* 391–401. doi:10.1016/j.system.2010.06.013.
[4] Arbuckle, J. L. (2006). Amos (Version 7.0.0) [Computer software]. Spring House, PA: Amos Development Corporation.
[5] Arbuckle, J. L., & Wothke, W. (1995). Amos 4.0 user's guide. Chicago: SmallWaters Corporation.
[6] Bentler, P. M. (2010). EQS (Version 6.1) for Windows [Computer software]. Encino, CA: Multivariate Software.
[7] Byrne, B. M. (2006). Structural equation modeling with EQS (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
[8] De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34,* 5–34. doi: 10.1017/S0272263111000489.
[9] De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn J. H. (in press). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics.* doi: http://dx.doi.org/10.1017/S0142716412000069.
[10] Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language. *Applied Linguistics, 21,* 354–375. doi: 10.1093/applin/21.3.354.
[11] Funato, S., & Ito, H. (2008). An empirical study on basic requirements for Japanese EFL learners to achieve oral fluency in English. *Annual Review of English Language Education in Japan, 19,* 41–50.
[12] Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal, 36,* 153–166. doi: 10.1080/09571730802389983.
[13] Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30,* 461–473. doi: 10.1093/applin/amp048.
[14] Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling, 6,* 1–55. doi: 10.1080/10705519909540118.
[15] Ishizuka, H. (2000). Goichishikino fukasato speakingnouryokuno soukan [Correlations between depth of vocabulary knowledge and speaking ability]. *STEP Bulletin, 12,* 13–25.
[16] Japan Association of College English Teachers. (JACET). Basic Word Revision Committee. (Ed.). (2003). *JACET List of 8000 Basic Words.* Tokyo: Author.
[17] Kline, R. B. (2010). Principles and practice of structural equation modeling (3rd ed.). New York: Guilford Press.
[18] Koizumi, R. (2005). Predicting speaking ability from vocabulary knowledge. *JLTA Journal, 7,* 1–20.
[19] Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures. *System, 40,* 522–532. doi: http://dx.doi.org/10.1016/j.system.2012.10.017.
[20] Kormos, J. (2006). Speech production and second language acquisition. Mahwah, NJ: Lawrence Erlbaum Associates.
[21] Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength. *Language Testing, 21,* 202–226.
[22] Levelt, W. J. M. (1989). Speaking. MA: MIT Press.
[23] Meara, P. (2005). Designing vocabulary tests for English, Spanish, and other languages. In C. S. Butler, M. Á. Gómez-González, & S. M. Doval-Suárez (Eds.), *The dynamics of language use* (pp. 271–285). Amsterdam, the Netherlands: John Benjamins.
[24] Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. D. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol, U.K: Multilingual Matters.
[25] Mochizuki, M., Murata, M., Uemura, T., Aizawa, K., Tono, Y., Sugimori, N., Shimizu, S. (2010). Ginobetsuoyobi sogotekieigoryokuwo suiteisuru goitestnokaihatsu [Development of a vocabulary test battery estimating English skills and proficiency]. Report of the Grant-in-Aid for Scientific Research (B)(2007-2009), Supported by Japan Society for the Promotion of Science. Project No. 19320084.
[26] Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620. doi: 10.1207/S15328007SEM0904_8.
[27] Muthén, L. K., & Muthén, B. O. (2011). Mplus (Version 6.12) [Computer software]. Los Angeles, CA: Muthén & Muthén.
[28] Pearson Education, Inc. (2008). Versant™ English Test.
[29] http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf (accessed 1/12/2012).
[30] Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance. *Language Learning, 52,* 513–536. doi: 10.1111/1467-9922.00193.
[31] Qian, D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge for assessing reading performance. *Language Testing, 21,* 28–52. doi:10.1191/0265532204lt273oa.
[32] Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10,* 355–371. doi:10.1177/026553229301000308.
[33] Schmitt, N. (2010). Researching vocabulary. New York: Palgrave MacMillan.
[34] Schmitt, N. (2012, March). Size and depth of vocabulary. Paper presented at the American Association of Applied Linguistics 2012, Boston, MA, U.S.
[35] Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing. *Language Learning, 53,* 165–202. doi: 10.1111/1467-9922.00213.
[36] Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition. *Studies in Second Language Acquisition, 26,* 173–199. doi: 10.1017/S0272263104262027.
[37] Skehan, P. (2009). Modeling second language performance. *Applied Linguistics, 30,* 510–532. doi: 10.1093/applin/amp047.
[38] Society for Testing English Proficiency (STEP). (2012). Investigating the relationship of the EIKEN tests with the CEFR. http://stepeiken.org/research (accessed 1/12/2012).

[39] Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31,* 577–607. doi:10.1017/S0272263109990039.
[40] Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Needham Heights, MA: Allyn & Bacon.
[41] Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 238–273). Amsterdam, the Netherlands: John Benjamins.
[42] Uenishi, K. (2006). Nihonjin eigogakushushano speakingryoku nikansuru jisshotekikenkyu [An empirical study regarding speaking ability of Japanese learners of English]. Eigo to eigo kyouiku tokubetsugo [English and English Language Education (Special ed.): Ozasa Toshiaki's retirement festschrift] (pp. 135–144). Hiroshima University.
[43] Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed and metacognitive knowledge in first and second language reading comprehension. *Journal of Educational Psychology, 96,* 19–30. doi: 10.1037/0022-0663.96.1.19.
[44] Van Moere, A. (2012). A Psycholinguistic Approach to Oral Language Assessment. *Language Testing, 29,* 325–344. doi: 10.1177/0265532211424478.
[45] Wood, D. (2010). Formulaic language and second language speech fluency. London, England: Continuum.

Rie Koizumi, Ph.D. (Linguistics, University of Tsukuba), is an Associate Professor at Juntendo University, Japan. She is interested in modeling factor structures of language ability and performance. Her website has supplementary materials for this article (http://www7b.biglobe.ne.jp/~koizumi/Koizumi_research.html).

Yo In'nami, Ph.D. (Linguistics, University of Tsukuba), is an Associate Professor of English at Shibaura Institute of Technology, Japan. He is interested in meta-analytic inquiry into the variability of effects and the longitudinal measurement of change in language proficiency.