THE JOURNAL OF ASIA TEFL
Vol. 17, No. 2, Summer 2020, 335-348
http://dx.doi.org/10.18823/asiatefl.2020.17.2.2.335

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511 © 2004 AsiaTEFL.org. All rights reserved.

Examining the Validity of the LexTALE Test for Japanese College Students

Tatsuya Nakata
Rikkyo University, Japan

Yu Tamura
Kansai University, Japan

Scott Aubrey
The Chinese University of Hong Kong, Hong Kong

The question of how vocabulary knowledge of second language (L2) learners can be measured in a valid and reliable way has attracted attention from researchers. One widely used format for assessing vocabulary knowledge is a Yes/No test, where learners are asked to indicate whether they know each vocabulary word on the test. The purpose of this study was to examine whether the LexTALE test, a recently developed Yes/No English vocabulary test, can be an approximate measure of vocabulary knowledge and general proficiency for Japanese learners of English. In this study, 111 Japanese university students majoring in English took the LexTALE, an English to Japanese translation test, and the Vocabulary Size Test (VST). They were further asked to provide self-ratings of their English proficiency. Analysis showed that the LexTALE score correlated more strongly with the translation score and VST score than self-ratings of their proficiency. The results also showed that the LexTALE score correlated significantly with the TOEFL ITP® score, although some self-ratings resulted in a higher correlation. The findings suggest that for Japanese learners of English, LexTALE may be used as an approximate measure of English vocabulary knowledge and, to a lesser extent, general proficiency.

Keywords: LexTALE, vocabulary size test, yes/no test, lexical decision task, translation test

Background

Many researchers have investigated the issue of how second language (L2) learners' vocabulary knowledge can be measured in a valid and reliable way. Debates around how to measure vocabulary knowledge are ongoing (e.g., Chujo & Oghigian, 2009; Enayat, Amirian, Zareian, & Ghaniabadi, 2018; Gyllstad, Vilkaitė, & Schmitt, 2015; Milton, 2009; Yue & Fan, 2016) and relate to the multifaceted nature of what it is to know a word. The two most prominent dimensions of vocabulary knowledge are the breadth and depth of word knowledge (Gyllstad, 2013). Vocabulary depth relates to the quality of word knowledge (i.e., how well words are known), whereas vocabulary breadth refers to the size of word knowledge, or the number of words a learner knows (Schmitt, 2014). Vocabulary size has received considerably more attention from researchers (David, 2008; Enayat et al., 2018; Lee & Kwon, 2014; Tang, 2007), owing possibly to the fact that knowing more words is an essential component of language knowledge (Nation, 2011) and strongly related to learners' language proficiency such as reading, listening, and writing skills (Koda, 1989; Laufer, 1992; Schmitt, Jiang, & Grabe, 2011; Stæhr, 2008). This has implications for the use of tests of vocabulary size, not just for researchers who may need to classify learners into proficiency levels, but for teachers who use these tests as a pedagogic tool as part of a placement test or a component of a diagnostic test (Thornbury, 2002), or to determine how well courses are meeting learning objectives (Beglar, 2010). Results from vocabulary tests, for instance, can be used to estimate students' general vocabulary knowledge, which could help the selection or development of class materials. Furthermore, considering that knowledge of vocabulary size is found to correlate significantly with general L2 proficiency (e.g., Harrington & Carey, 2009; Meara & Jones, 1988; Shillaw, 1996), vocabulary size tests may be used as a low-cost alternative to standardized proficiency tests such as TOEIC® or TOEFL®. The current research aims to examine whether the LexTALE test, a recently developed test of vocabulary size, can be a valid measure of vocabulary knowledge and general proficiency for Japanese learners of English.

In EFL (English as a Foreign Language) contexts, such as Japan, assessing English vocabulary knowledge has unique challenges. In such acquisition-poor environments, where students rarely communicate in English outside of the classroom, vocabulary learning often takes the form of memorizing translated definitions of discrete, decontextualized word lists (Gu, 2003). Japanese high school students, for example, are required to learn such large quantities of English words that their depth of understanding of each word is often very shallow (Ruegg, 2007), so the concept of knowing a word may be very different than if learners had regular exposure to the target language. Moreover, cultural tendencies may affect how learners interact with a vocabulary test. As Japanese tend to maintain a more self-critical attitude than their western counterparts (Heine, Kitayama, & Lehman, 2001), self-report vocabulary tests, in which learners are required to judge their certainty of knowing a word, could have external reliability problems – that is, in comparison with other populations, Japanese learners may under-report the number of words they know. Yet, estimating vocabulary size may be especially important in contexts like Japan, where there is a reliance on standardized tests, such as TOEIC® and TOEFL®, for placement and assessment, which are both expensive and time-consuming. A valid but simple vocabulary knowledge test could provide an accessible alternative for some of these purposes.

The most widely used format for assessing vocabulary size is perhaps the multiple-choice format, where learners are asked to choose the most appropriate meaning corresponding to an L2 word, or vice versa. Examples of multiple-choice vocabulary tests include the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001) and the Vocabulary Size Test (hereafter, VST; Nation & Beglar, 2007). Relevant to the current study is the VST, which was designed to measure a learner's receptive vocabulary size and consists of 140 multiple-choice items, with 10 items from each 1,000 word family level (Nation & Beglar, 2007). Beglar (2010) conducted a validation study of the VST with 178 Japanese undergraduate and postgraduate students, a population very similar to that of the present study. He analyzed VST scores in terms of content, substantive and structural generalizability, as well as responsiveness and interpretability. The results suggested that the VST may be a valid and reliable measure of written receptive vocabulary size. Since then, bilingual versions of the VST have been developed (Japanese, Korean, Vietnamese, Mandarin, Russian, and Persian) which have increased its prominence as a vocabulary measure in L2 vocabulary research (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011; Zhang, 2013). At the same time, recent research also suggests that due to the guessing problem (i.e. the possibility that test-takers can guess the correct answer regardless of their knowledge or ability), the VST tends to overestimate vocabulary knowledge. Instead, some researchers consider a meaning recall test (where participants are asked to provide definitions or L1 translations of L2 words) to be a more valid measure (Gyllstad, 2013; Gyllstad et al., 2015; Stewart, 2014; Stoeckel et al., 2019; Stoeckel, Bennett, & McLean, 2016).

As an alternative to multiple-choice tests, another common format for assessing vocabulary knowledge is a Yes/No test, also referred to as a checklist test (e.g., Huibregtse, Admiraal, & Meara, 2002; Lemhöfer & Broersma, 2012; McLean, Stewart, & Batty, 2020; Meara & Jones, 1988; Stubbe, 2012; Stubbe &

Stewart, 2012). In this test, learners are presented with L2 words and asked to indicate whether they know each word. To discourage participants from claiming that they "know" all words in the test, non-words (words that do not exist in the L2) are usually included in the test, and scores are adjusted when learners claim to know any non-words. However, some researchers recommend the use of real low-frequency words over non-words because adjusting scores based on learners' responses to non-words often results in the underestimation of vocabulary knowledge, especially for higher proficiency learners (Stubbe, 2012). It is argued that using real low-frequency words, rather than non-words, will also be a better use of time, as well as increase the face validity of tests (e.g., Mochida & Harrington, 2006; Stubbe, 2012). In a variation of a Yes/No test known as a lexical decision task (e.g., Lemhöfer & Broersma, 2012), learners are presented with a series of letter strings, and asked to indicate whether they are real words in the L2, regardless of whether learners know them. The Yes/No test format has two main advantages over a multiple-choice test. One advantage is that a Yes/No test is quicker to administer. This means that more items can be tested in the same amount of time, which may help increase reliability of the test scores (Nation, 2013). The second advantage is that it is easier for teachers, administrators or researchers to grade, which may lead to more widespread use than other tests (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001).

Researchers have investigated the validity of Yes/No tests as a measure of L2 vocabulary knowledge. Mochida and Harrington (2006), for instance, examined the relationship between scores on a Yes/No test and the Vocabulary Levels Test (Schmitt et al., 2001) with ESL undergraduate and postgraduate students in Australia. They report that the correlation coefficients (r) between the scores on the two tests were between .85 and .88, indicating a close relationship between performance on a Yes/No test and receptive vocabulary knowledge. In a study conducted by Eyckmans (2004), French-speaking students at a university in Belgium took a Yes/No test of Dutch, which consisted of 60 words and 40 non-words. One week later, the participants were asked to translate the same 60 words used on the Yes/No test into French. Eyckmans reports a significant correlation ($.39 \leq r \leq .60$) between scores on the Yes/No test and the translation test, indicating a moderate relationship between performance on the two tests. For Japanese college students, Barrow, Nakanishi, and Ishino (1999) found a significant correlation ($r = .73$) between performance on a Yes/No test and L2 (English) to L1 (Japanese) translation test while Stubbe (2012) reported a very high correlation ($.80 \leq r \leq .83$) between performance on a Yes/No test and a multiple-choice test. In contrast to these findings, Cameron (2002) failed to find any statistically significant correlation between performance on a Yes/No test and Vocabulary Levels Test with 15-year-old ESL students in the UK – though the results of her study may be due in part to the very small sample size ($n = 14$). Overall, these findings suggest that a Yes/No test may be a good predictor of L2 vocabulary size.

Some researchers also report a significant correlation between performance on a Yes/No test and general L2 proficiency. Meara and Jones (1988), for instance, found a relatively strong correlation ($.67 \leq r \leq .71$) between scores on a Yes/No test and a placement test (the Joint Entrance Test) at an English language school in the UK. Shillaw (1996) administered an English Yes/No test to Japanese university students. He found a significant correlation ($.42 \leq r \leq .48$) between scores on the Yes/No test and a general proficiency test measuring listening, vocabulary, reading comprehension, and grammar knowledge. Similarly, Harrington and Carey (2009) found a significant correlation ($.54 \leq r \leq .64$) between the performance of ESL students in Australia on a Yes/No test and the placement level decisions at a language school based on the results of a proficiency test measuring listening, grammar, writing, and speaking skills. These findings suggest that a Yes/No test may be a good predictor of not only vocabulary knowledge but also general proficiency in an L2.

LexTALE: A Recently Developed and Validated Yes/No Test

One particular type of Yes/No test that has been increasingly used by researchers in recent years is the LexTALE (Lexical Test for Advanced Learners of English; Lemhöfer & Broersma, 2012). The test was originally developed in the field of psycholinguistics. Psycholinguists have been interested in how various factors, such as synonymy, homonymy, or cognacy, affect L2 processing. When conducting experiments with L2 learners, it is crucial to control participants' proficiency levels, especially vocabulary knowledge. The majority of existing studies, however, have relied solely on self-ratings or questionnaires to estimate L2 proficiency or vocabulary knowledge. LexTALE was developed to help provide researchers with a quick and approximate estimate of vocabulary knowledge, and possibly general proficiency. One potential advantage of LexTALE lies in its convenience; since the test requires only 3.5 to 5 minutes to complete (Lemhöfer & Broersma, 2012), it has the potential to be used as a standard proficiency measure in research.

In LexTALE, participants are presented with 60 stimuli and asked to indicate whether each stimulus is an English word. The stimuli consist of 40 words and 20 non-words taken from an unpublished 10K test developed by Meara (1996; cited in Lemhöfer & Broersma, 2012). Lemhöfer and Broersma (2012) conducted a validation study to investigate whether LexTALE could be a good predictor of vocabulary knowledge and general proficiency in English. They found a significant correlation between performance on LexTALE and translation test scores for Dutch ($.59 \leq r \leq .78$) and Korean ($.41 \leq r \leq .51$) learners of English. They also report that LexTALE scores correlate significantly with scores on two English proficiency tests: the Quick Placement Test (Dutch: $.58 \leq r \leq .63$; Korean: $.29 \leq r \leq .30$; University of Cambridge Local Examination Syndicate, 2001) and TOEIC® Test (Korean: $.33 \leq r \leq .35$). Furthermore, Lemhöfer and Broersma found that the LexTALE score was a better predictor of translation test scores than self-ratings of proficiency provided by participants, although self-ratings sometimes produced a higher correlation with a measure of general proficiency (the Quick Placement Test) than LexTALE. Considering that LexTALE only takes 3.5 to 5 minutes to administer, Lemhöfer and Broersma argue that the test can be used as a quick measure of English vocabulary knowledge and, to a lesser extent, general proficiency. Previous research also suggested that the LexTALE score correlated more strongly with performance on an online lexical decision task (Lemhöfer & Dijkstra, 2004) and progressive demasking (PDM) task (Lemhöfer et al., 2008) than self-ratings. These findings suggest that the LexTALE score may be a good predictor of not only explicit vocabulary knowledge measured by paper-and pencil tests, but also proceduralized knowledge measured by online tasks performed under time pressure.

One important issue when scoring a Yes/No test such as the LexTALE is how to correct for guessing. The most straightforward way to calculate the score would be to simply calculate the proportion of correct responses (i.e., responding *yes* to real words and *no* to non-words). Another scoring procedure often used by researchers is the Δm vaule proposed by Meara (1992 cited in Lemhöfer & Broersma, 2012), where false alarms (i.e., responding *yes* to non-words) are penalized. Δm is calculated using the following formula (Huibregtse et al., 2002):

$$\Delta m = (\mathrm{h} - \mathrm{f}) / (1 - \mathrm{f}) - \mathrm{f} / \mathrm{h}$$

Note. In the above formula, *h* refers to a hit rate (proportion of responding *yes* to real words), and *f* refers to a false alarm rate (proportion of responding *yes* to non-words).

Huibregtse et al. (2002) also proposed a scoring procedure known as $\mathrm{I_{SDT}}$. $\mathrm{I_{SDT}}$ not only corrects for guessing but also takes response bias of each participant into account (e.g., when they are not sure, some participants are likely to choose *yes*, whereas others are more conservative). $\mathrm{I_{SDT}}$ can be calculated using the following formula (where *h* refers to a hit rate, and *f* refers to a false alarm rate):

$$\mathrm{I_{SDT}} = 1 - (4\mathrm{h}\,(1 - \mathrm{f}) - 2\,(\mathrm{h} - \mathrm{f})\,(1 + \mathrm{h} - \mathrm{f})) / (4\mathrm{h}\,(1 - \mathrm{f}) - (\mathrm{h} - \mathrm{f})\,(1 + \mathrm{h} - \mathrm{f}))$$

In a validation study of the LexTALE, Lemhöfer and Broersma (2012) administered the test to Dutch and Korean learners of English, and compared the following three scoring protocols: %Correctav, Δm, and $\mathrm{I_{SDT}}$. %Correctav refers to "averaged % correct," and is calculated by averaging the proportion of correct responses for words and non-words; because LexTALE consists of 40 words and 20 non-words, simply calculating the average of all items means that the tendency to reject real words is penalized more than the tendency to accept non-words. %Correctav takes into account the unbalanced proportion of words and non-words in LexTALE by averaging % correct for words and non-words. Lemhöfer and Broersma found that the three scoring procedures (%Correctav, Δm, and $\mathrm{I}_{\mathrm{SDT}}$) produced somewhat similar results. The findings are consistent with earlier studies suggesting that the scoring procedure has little effect on Yes/No test scores (e.g., Mochida & Harrington, 2006; Stubbe, 2012).

The Present Study

Previous research suggests that LexTALE can be used as an approximate measure of English vocabulary knowledge and, possibly, general proficiency. At the same time, the existing studies have some limitations. First, although the validity of LexTALE has been tested with Dutch and Korean learners of English, it is not clear whether the test can be used as a measure of English vocabulary knowledge for other populations. Considering that learners from different cultures may behave differently when taking a Yes/No test (Lemhöfer & Broersma, 2012), it is important to investigate the validity of the test with other populations. The present study examined whether LexTALE scores can be used as a predictor of vocabulary knowledge and general proficiency for Japanese learners of English. Second, Lemhöfer and Broersma (2012) examined whether LexTALE scores can be a better predictor of vocabulary knowledge than self-ratings of proficiency by participants. However, their self-ratings were only concerned with estimating each of the four skills of reading, writing, speaking, and listening. If the researchers had also asked learners for self-ratings of vocabulary knowledge, the learners' self-ratings might have been more useful as predictors of vocabulary knowledge. Self-ratings in this study, therefore, included those of vocabulary knowledge.

With potential limitations of the existing studies in mind, the present study aimed to examine whether LexTALE is a better predictor of English vocabulary knowledge and general English proficiency than self-ratings for Japanese college students. The research questions (RQs) of the current study are as follows:

RQ 1: Is LexTALE a better predictor of English vocabulary knowledge than self-ratings for Japanese college students?
RQ 2: Is LexTALE a better predictor of general English proficiency than self-ratings for Japanese college students?
RQ 3: Which scoring procedure for LexTALE will provide the most accurate assessment of English vocabulary knowledge and general English proficiency for Japanese college students?

To examine the above three RQs, Japanese university students majoring in English took the LexTALE, an English to Japanese translation test, and the VST. Scores on the translation test and the VST were used as criterion measures of vocabulary knowledge. A translation test was chosen as a criterion measure because previous research suggests that it provides a more accurate assessment of receptive vocabulary knowledge than a multiple-choice test (e.g., Stoeckel et al., 2019). The VST was also used as another criterion measure because despite its tendency to overestimate vocabulary knowledge due to the guessing problem, a multiple-choice test is among the most widely used formats for assessing vocabulary knowledge in existing research (e.g., Beglar, 2010; Elgort, 2013; Gyllstad et al., 2015; Karami, 2012; Nguyen & Nation, 2011; Zhang, 2013). Scores on the TOEFL ITP® (which consists of listening, reading, and structure and written expression sections) were used as a criterion measure of general proficiency. Participants were further asked to provide self-ratings of their English proficiency to examine whether LexTALE would result in a better predictor of English vocabulary knowledge (RQ 1) and general English proficiency (RQ 2) than self-ratings. To examine the effects of different scoring procedures (RQ 3), responses on the LexTALE were scored using the following four systems: %Correct, %Correctav, Δm, and $\mathrm{I}_{\mathrm{SDT}}$ (see the Method section for details).

Method

The Present Study

The participants were 111 first-year English majors at a private university in Japan whose L1 was Japanese. Their average TOEFL IPT® score was 502.90 (*SD* = 32.50), which is estimated to fall between B1 and B2 levels in the Common European Framework of Reference for Languages (CEFR) benchmark (Educational Testing Service, 2019). Their English proficiency was estimated to be lower than that of the Korean and Dutch learners of English in Lemhöfer and Broersma's (2012) study because (a) the Korean participants in the earlier study had an average TOEIC® score of 887, which is estimated to fall between B2 and C1 levels in the CEFR (Educational Testing Service, 2015), and (b) Dutch participants in the earlier study had a higher score (76.8%) than their Korean counterparts (64.1%) on a measure of general English proficiency (the Quick Placement Test).

Procedure

Data were collected over two regular class sessions. In the first session, participants took the three tests in the following order: the LexTALE, the translation test, and the VST (Nation & Beglar, 2007). All three tests were administered online. The LexTALE test administered in this study was exactly the same as the one used in Lemhöfer and Broersma (2012) except that the spelling for one item (*savoury*) was changed to the American spelling (*savory*), since American spelling is typically taught in Japanese high schools and universities. In the LexTALE test, participants were presented with 63 stimuli (the first three were practice items and were not scored). They were asked to indicate whether each stimulus was an English word by clicking "Yes" for a word or "No" for a non-word. The 60 critical stimuli consisted of 40 words and 20 non-words. The 40 words were from a wide range of frequency levels, from the most frequent 1,000 word families to beyond the 10,000 word level. The breakdown of the frequency levels of the target words is shown in Table 1. In the translation test, participants were asked to translate 30 English words into their L1 (Japanese). The same 30 words from the study conducted by Lemhöfer and Broersma (2012) were used. The items in the translation test also consisted of those from a wide range of frequency levels, as shown in Table 1. After the translation test, a bilingual Japanese version of the VST (Sasao & Nakata, 2010) was administered.

TABLE 1
Frequency Levels of Stimuli Used in the LexTALE and Translation Test

Frequency level	LexTALE		Translation Test	
	n	%	*n*	%
1,000	1	3%	0	0%
2,000	3	8%	6	15%
3,000	2	5%	4	10%
4,000	7	18%	7	18%
5,000	4	10%	4	10%
6,000	8	20%	3	8%
7,000	5	13%	3	8%
8,000	2	5%	2	5%
9,000	0	0%	0	0%
10,000	5	13%	1	3%
Beyond 10,000	3	8%	0	0%

Note. Frequency levels were calculated using Compleat Lexical Tutor (Cobb, n.d.) using BNC-COCA lists.

Three weeks after the first session, participants were asked to provide self-ratings of their English proficiency on a 100-point scale, anchored by 1 (Very low) and 100 (Very high). Participants were asked to evaluate their English proficiency in the following areas: reading, writing, listening, speaking, grammar, vocabulary, and pronunciation. The self-ratings were administered 3 weeks after the first session, instead of on the same day, because asking self-ratings immediately after giving relatively demanding tests (i.e., the LexTALE, the translation test, and the VST) might influence their self-efficacy, potentially affecting self-ratings of their proficiency.

Scoring

Scores on the LexTALE were calculated using the following four scoring procedures: %Correct, %Correctav, Δm, and I_{SDT}. %Correct was calculated by averaging the proportion of correct responses for all test items. %Correctav was calculated by averaging the proportion of correct responses for words and non-words. Δm and I_{SDT} were calculated using the same formulas described in the Background section. For the translation test, to maintain consistency in scoring, answer keys (a list of correct responses) were first created by two of the authors based on English-Japanese dictionaries. Responses on the translation test were then categorized by custom software into the following three groups based on the answer keys: (a) correct (i.e., the response matches one of the answer keys for the target word), (b) blank responses, and (c) other responses. The responses that were categorized as (c), which accounted for 24% of all responses, were independently scored by two of the authors. The inter-rater agreement was 98.6% for the responses in category (c). For the VST, the estimated vocabulary size of the participants was calculated by multiplying the number of correct responses by 100.

Method

Table 2 provides descriptive statistics for the LexTALE, the translation test, the VST, TOEFL ITP®, and self-ratings. For the LexTALE, the average hit rate (64%) was lower than those of Dutch (68%) and Korean (73%) learners in Lemhöfer and Broersma (2012). The average false alarm rate (45%) was higher than that reported for Dutch learners (17%), but similar to the one reported for Korean learners (42%).

TABLE 2
Descriptive Statistics for the LexTALE, Translation Test, Vocabulary Size Test, TOEFL ITP®, and Self-ratings

		M	*SD*	Min	Max
LexTALE	Hit rate	64%	15%	15%	98%
	False alarm rate	45%	22%	0%	95%
	%Correct	61%	8%	42%	92%
	%Correctav	59%	8%	39%	94%
	Δm	−0.39	0.55	−2.33	0.88
	I_{SDT}	0.21	0.18	−0.25	0.88
Translation Test		33%	11%	10%	67%
Vocabulary Size Test		8136.04	765.96	5700	10500
TOEFL ITP®		502.90	32.50	450	633
Self-ratings	Reading	50.57	23.15	10	100
	Writing	51.14	16.24	10	90
	Listening	53.43	20.08	10	100
	Speaking	45.74	21.36	5	100
	Grammar	52.36	17.98	10	90
	Vocabulary	51.63	17.46	10	90
	Pronunciation	51.14	23.18	5	100
	Total	356.00	115.00	70	660

Note. %Correct was calculated by averaging the proportion of correct responses for all test items. %Correctav was calculated by averaging the proportion of correct responses for words and non-words (see the Background section).

To examine relationships among the LexTALE scores, the translation test scores, the VST scores, the TOEFL ITP® scores, and the self-ratings, Pearson correlation coefficients (*r*) were calculated. Table 3 summarizes correlation coefficients. Scatterplots for key variables are provided in Figure 1 and 2.

TABLE 3
Correlation Coefficients (r) among the LexTALE, Translation Test, Vocabulary Size Test, TOEFL ITP®, and Self-ratings

		Translation	Vocabulary Size Test	TOEFL ITP®
LexTALE	%Correct	.46*** [.29, .59]	.43*** [.26, .57]	.37*** [.19, .51]
	%Correct$_{av}$	.45*** [.28, .58]	.38*** [.20, .52]	.43*** [.26, .57]
	Δ*m*	.38*** [.21, .53]	.31*** [.14, .47]	.35*** [.17, .50]
	I$_{SDT}$	.41*** [.24, .56]	.35*** [.18, .50]	.41*** [.24, .55]
Self-ratings	Reading	.19* [.01, .37]	.11 [−.08, .29]	.17 [−.01, .35]
	Writing	.19* [.00, .36]	.20* [.02, .38]	.26** [.08, .42]
	Listening	.19* [.00, .36]	.11 [−.08, .29]	.39*** [.22, .54]
	Speaking	.17 [−.02, .34]	.05 [−.14, .23]	.36*** [.18, .51]
	Grammar	.21* [.02, .38]	.06 [−.13, .24]	.31*** [.14, .48]
	Vocabulary	.27** [.09, .43]	.21* [.02, .38]	.23* [.05, .40]
	Pronunciation	.17 [−.02, .35]	.06 [−.12, .24]	.35*** [.17, .50]
	Total	.24* [.06, .41]	.13 [−.06, .31]	.36*** [.19, .51]
Translation Test		–	.51*** [.36, .64]	.55*** [.40, .66]
Vocabulary Size Test		–	–	.30** [.12, .46]

Note. *** $p < .001$. ** $p < .01$. * $p < .05$. 95% confidence intervals are shown in brackets.

Table 3 shows that the self-rating for vocabulary knowledge correlated more strongly with the translation scores and the VST scores than other self-ratings. The LexTALE scores, however, correlated more strongly with the translation scores and VST scores than any of the self-ratings. At the same time, the 95% confidence intervals for all correlation coefficients overlapped with each other. Table 3 also shows that for the TOEFL ITP® score, %Correct$_{av}$ and I$_{SDT}$ had a higher correlation than self-ratings. Some self-ratings, however, produced correlation coefficients that were no lower than %Correct or Δ*m*. The 95% confidence intervals for all correlation coefficients, once again, overlapped with each other for the TOEFL ITP® score.

Table 3 also shows that the correlation coefficients between the LexTALE and three criterion measures were similar, regardless of the scoring procedures: $.38 \leq r \leq .46$ for the translation test: $.31 \leq r \leq .43$ for VST and $.35 \leq r \leq .43$ for TOEFL ITP®. As shown in Figure 1, the scores produced by the four scoring procedures correlated significantly with each other ($.71 \leq r \leq .98$). In particular, the correlation coefficients among %Correct$_{av}$, Δ*m*, and I$_{SDT}$ were very high ($.95 \leq r \leq .98$).

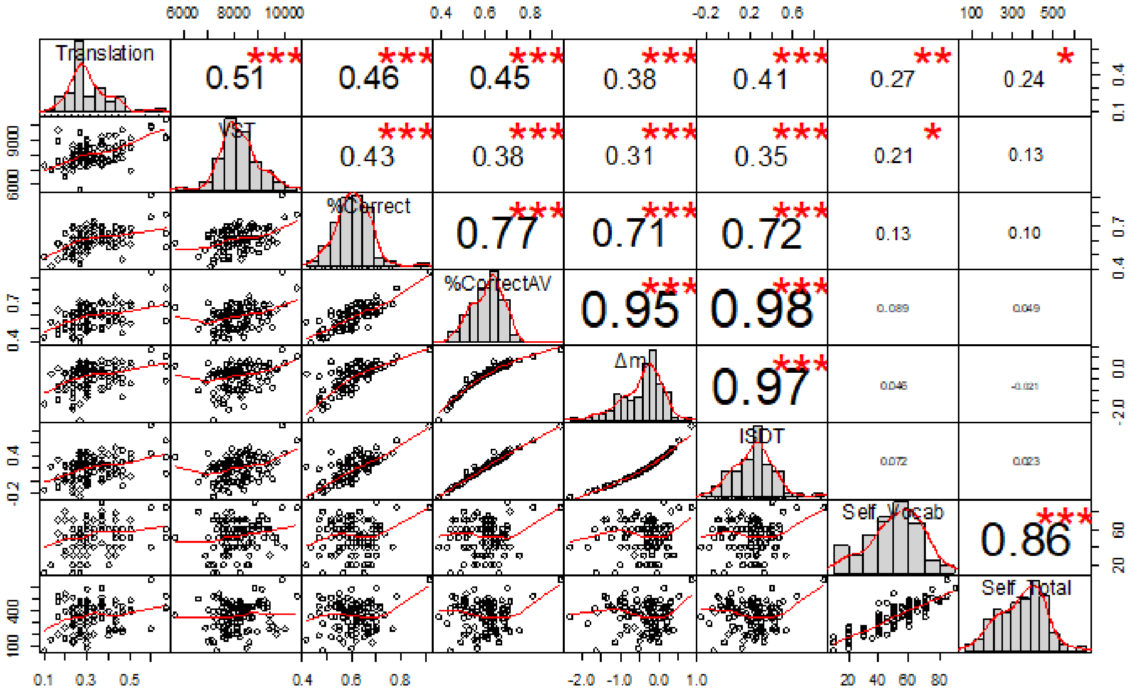

Figure 1. Scatterplots and correlation matrix for the LexTALE, translation test, Vocabulary Size Test (VST), and self-ratings.

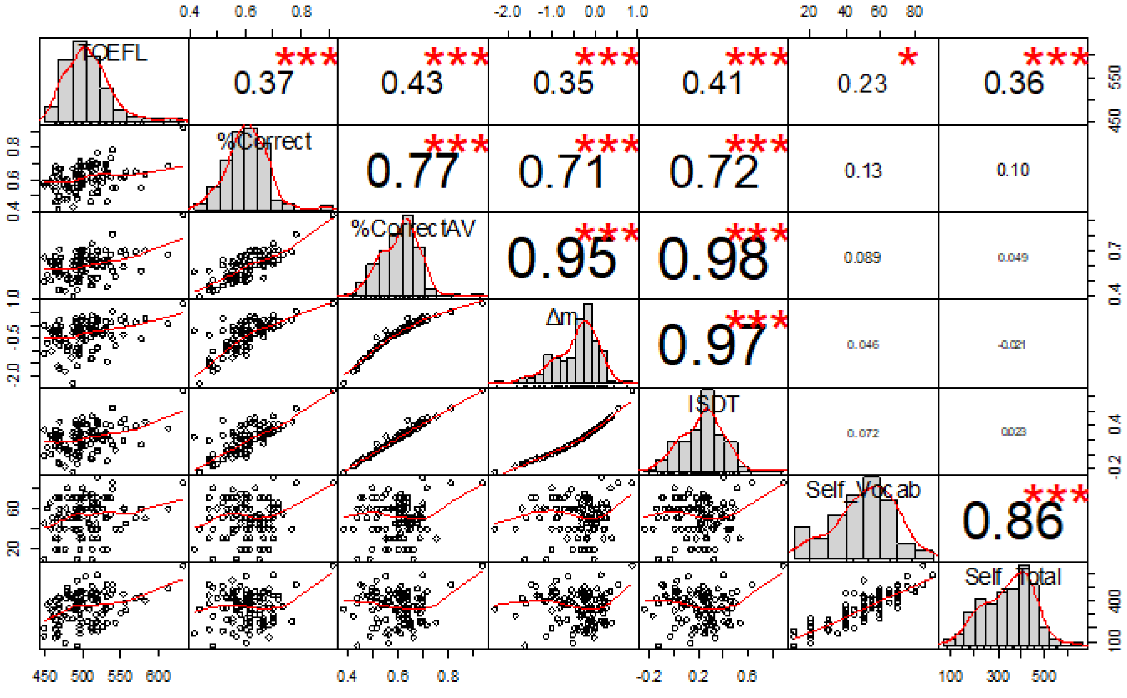

Figure 2. Scatterplots and correlation matrix for the LexTALE, TOEFL ITP®, and self-ratings.

Discussion and Conclusions

RQ 1 of this study asked whether LexTALE is a more accurate predictor of English vocabulary knowledge than self-ratings for Japanese learners of English. The present study showed that the LexTALE score correlated more strongly with the translation score ($.38 \le r \le .46$) and VST score ($.31 \le r \le .43$) than self-ratings ($.05 \le r \le .27$). The results suggest that the LexTALE score may be a better predictor of vocabulary knowledge than self-ratings. However, the results also showed that the 95% confidence intervals for all correlation coefficients overlapped with each other. Furthermore, although LexTALE had a higher correlation with the translation and VST scores than self-ratings, the correlation coefficients (*r*) between LexTALE and these two criterion measures were between .31 and 46, which are considered as having no more than a medium effect (Cohen, 1988). The results suggest that although LexTALE may be used as a quick and rough estimate of vocabulary knowledge, it is advisable to administer other vocabulary tests such as the translation test or VST when time allows, and when a more accurate assessment is needed.

Another purpose of this study was to investigate whether LexTALE may be used as a measure of general English proficiency (RQ 2). This study suggested that although %Correct$_{av}$ and I$_{SDT}$ had a higher correlation with the TOEFL ITP® score than self-ratings, %Correct and Δm failed to outperform some of the self-ratings. The findings are consistent with those reported for Dutch and Korean learners (Lemhöfer & Broersma, 2012), where self-ratings sometimes produced a higher correlation with a measure of general proficiency (Quick Placement Test) than LexTALE. The findings suggest that although the LexTALE scores may correlate significantly with a global measure of English proficiency, it may be more useful as a predictor of vocabulary knowledge than that of general proficiency.

The correlation between the LexTALE scores and vocabulary knowledge ($.29 \le r \le .38$) found in this study was lower than those reported for Dutch ($.59 \le r \le .78$) and Korean ($.41 \le r \le .51$) learners in Lemhöfer and Broersma (2012). This could be in part due to the proficiency differences between participants in the study and those in the Lemhöfer and Broersma's study. As described in the Participants section, the Dutch and Korean participants in Lemhöfer and Broersma (2012) were likely more advanced learners of English (approximately CEFR: B2-C1) than those in this study (approximately CEFR: B1-B2). Considering the relatively low proficiency of the participants, including more higher frequency words as target items might have resulted in better discrimination among participants.

RQ 3 of this study was concerned with effects of different procedures for scoring LexTALE. The present study showed that the four scoring procedures (%Correct, %Correct$_{av,}$ $\Delta m,$ I$_{SDT}$) produced very similar results ($.71 \le r \le .98$). The findings are consistent with those reported by Lemhöfer and Broersma (2012), as well as earlier research on the Yes/No test (e.g., Mochida & Harrington, 2006; Stubbe, 2012). Considering that %Correct and %Correct$_{av}$ are easier to calculate than Δm and I$_{SDT}$, %Correct and %Correct$_{av}$ may be used as an approximate measure for Δm and I$_{SDT}$. The findings are especially useful for instructors who may not have time or resources to calculate Δm or I$_{SDT}$.

Although LexTALE was designed for L2 researchers to measure their participants' vocabulary knowledge, the findings of this study also have pedagogical implications. Teachers who have students of intermediate proficiency (i.e., CEFR B1 or B2) could use LexTALE if they need a rough approximation of their students' vocabulary knowledge. LexTALE should be sufficient for diagnosing students' general word knowledge, which could help select or develop class materials. The usefulness of the LexTALE in comparison to other vocabulary tests, such as the translation test or the VST, lies in its convenience. While the VST takes approximately 40 minutes to complete (Nation, 2012), the LexTALE takes just 3.5 to 5 minutes (Lemhöfer & Broersma, 2012). Thus, the test can be administered with minimal interruption to classroom activities. We cannot recommend the LexTALE as a replacement for standardized proficiency tests in most cases, especially for high-stakes assessment, such as class placement, graduation, student exchange, or scholarship programs. However, considering the high costs associated with standardized tests such as TOEIC® or TOEFL® (i.e., the test itself, administrators, test venue, time), the LexTALE might be appropriate in low-stakes proficiency assessment where the focus is on raising students' awareness of one aspect of general proficiency. It could also be used as an informal formative assessment to track learners' progress over time if a goal of the course is to increase vocabulary size (provided that learners are not exposed to the same stimuli in different test administrations).

Despite these tentative conclusions, some limitations of the present study need to be highlighted. One limitation is that although the participants in this study have relatively high English proficiency for Japanese college students, their proficiency level might have still been lower than the intended target population for the LexTALE test. Examining the validity of LexTALE with Japanese learners with even higher proficiency would be a useful follow-up to this study. Another limitation is that an external criterion measure of vocabulary knowledge used (the translation test and the VST) only measured explicit knowledge. Given the importance of proceduralized knowledge for fluent language processing, in future research, it would be useful to examine a relationship between the LexTALE scores and performance on online measures such as a timed lexical decision task (Lemhöfer & Dijkstra, 2004) or progressive demasking (PDM) task (Lemhöfer et al., 2008). Considering that earlier research has indicated the potential of LexTALE as a quick measure of English vocabulary knowledge, further research investigating the validity of the test is warranted.

Acknowledgements

This is a product of research which was financially supported in part by the Kansai University Fund for Supporting Young Scholars, 2018. "What vocabulary factors are crucial for the assessment and development of successful second language speech?" The authors are grateful to Atsushi Mizumoto for his cooperation with data collection.

The Authors

Tatsuya Nakata is Associate Professor at the College of Intercultural Communication, Rikkyo University, Japan. His research interests include L2 vocabulary acquisition and computer-assisted language learning. His research has appeared in publications such as *Studies in Second Language Acquisition, Modern Language Journal, TESOL Quarterly, Language Teaching Research, and Second Language Research.*

Department of Intercultural Communication
College of Intercultural Communication
Rikkyo University
3-34-1 Nishi-Ikebukuro, Toshima-ku, Tokyo 171-8501 Japan
Tel: +81 3-3985-2231
Email: nakata@rikkyo.ac.jp

Yu Tamura received his PhD from Nagoya University, Japan. He is currently an associate professor in the Faculty of Foreign Language Studies at Kansai University.

Scott Aubrey is an Assistant Professor in the Department of Curriculum and Instruction, Faculty of Education, at the Chinese University of Hong Kong, where he teaches English language teaching methodology courses to BA/BEd and MA students. His research interests include L2 motivation, task-based language teaching, and L2 writing instruction.

References

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27*, 223–247. doi:10.1016/S0346-251X(99)00018-4

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*, 235–274. doi:10.1177/026553220101800301

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*, 101–118. doi:10.1177/0265532209340194

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*, 145–173.

Cobb, T. (n.d.). The Compleat Lexical Tutor. Retrieved from https://www.lextutor.ca/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Chujo, K., & Oghigian, K. (2009). How many words do you need to know to understand TOEIC, TOEFL & EIKAN? An examination of the text coverage and high frequency vocabulary. *The Journal of Asian TEFL, 6*(2), 121–148.

David, A. (2008). Vocabulary breadth in French L2 learners. *The Language Learning Journal*, *36*, 167–180. doi:10.1080/09571730802389991

Educational Testing Service. (2015). Mapping the TOEIC tests on the CEFR. Retrieved July 14, 2019, from https://www.etsglobal.org/Global/Eng/content/download/890/13127/version/9/file/Mapping+the+TOEIC+and+TOEIC+Bridge+Tests+on+the+CEFR+-+MAR338-LR.pdf

Educational Testing Service. (2019). TOEFL: For academic institutions: Compare scores. Retrieved March 14, 2019, from https://www.ets.org/toefl/institutions/scores/compare/

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*, 253–272. doi:10.1177/0265532212459028

Enayat, M. J., Amirian, S. M. R., Zareian, G., & Ghaniabadi, S. (2018). Reliable measure of written receptive vocabulary size: Using the L2 depth of vocabulary knowledge as a yardstick. *SAGE Open*, *8*, 1–15. doi:10.1177/2158244017752221

Eyckmans, J. (2004). *Measuring receptive vocabulary size: Reliability and validity of the yes/no vocabulary test for French-speaking learners of Dutch* (Unpublished doctoral dissertation). Utrecht University, Utrecht, Netherlands.

Gu, P. Y. (2003). Brush and freehand: The vocabulary-learning art of two successful Chinese EFL learners. *TESOL Quarterly*, *37*, 73–104. doi:10.2307/3588466

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective—Challenges and potential solutions. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 11–28). Amsterdam, Netherlands: European Second Language Association.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL – International Journal of Applied Linguistics*, *166*, 278–306. doi:10.1075/itl.166.2.04gyl

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37*, 614–626. doi:10.1016/j.system.2009.09.006

Heine, S. J., Kitayama, S., & Lehman, D. R. (2001). Cultural differences in self-evaluation: Japanese readily accept negative self-relevant information. *Journal of Cross-Cultural Psychology*, *32*, 434–443. doi:10.1177/0022022101032004004

Huibregtse, I., Admiraal, W., & Meara, P. M. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*, 227–245. doi:10.1191/0265532202lt229oa

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, *43*, 53–67. doi:10.1177/0033688212439359

Koda, K. (1989). The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Language Annals*, *22*, 529–540. doi:10.1111/j.1944-9720.1989.tb02780.x

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London, UK: Macmillan.

Lee, K., & Kwon, S. (2014). Effects of vocabulary memorizing tools on learners' vocabulary size. *The Journal of Asia TEFL, 11*(2), 125–148.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. doi:10.3758/s13428-011-0146-0

Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, *32*, 533–550. doi:10.3758/BF03195845

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*, 12–31. doi:10.1037/0278-7393.34.1.12

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*. doi:10.1177/0265532219898380

Meara, P. M. (1992). *New approaches to testing vocabulary knowledge* (Unpublished manuscript). Swansea, UK: Centre for Applied Language Studies, University of Wales.

Meara, P. M. (1996). *English Vocabulary Tests: 10k* (Unpublished manuscript). Swansea, UK: Centre for Applied Language Studies, University of Wales.

Meara, P. M., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80–87). London, UK: CILT.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Mochida, K., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*, 73–98. doi:10.1191/0265532206lt321oa

Nation, I. S. P. (2011). Second language speaking. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 444–454). New York, NY: Routledge.

Nation, I. S. P. (2012). The vocabulary size test: Information and specifications. Retrieved July 14, 2019, from https://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*, 86–99. doi:10.1177/0033688210390264

Ruegg, R. (2007). The English vocabulary level of Japanese junior high school students. *JALT 2007 Conference Proceedings*, 103–109.

Sasao, Y., & Nakata, T. (2010). Vocabulary Size Test (bilingual Japanese version). Retrieved from https://www.wgtn.ac.nz/lals/about/staff/Publications/paul-nation/Vocab_Size_Test_Japanese.pdf

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*, 913–951. doi:10.1111/lang.12077

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43. doi:10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*, 55–89.

Shillaw, J. (1996). The application of Rasch modelling to yes/no vocabulary tests. *The Lognostics Virtual Library*. Retrieved July 14, 2019, from http://www.lognostics.co.uk/vlibrary/shillaw1996.doc

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*, 139–152. doi:10.1080/09571730802389975

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282. doi:10.1080/15434303.2014.922977

Stoeckel, T., Bennett, P., & McLean, S. (2016). Is "I Don't Know" a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly*, *50*, 965–975. doi:10.1002/tesq.325

Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning recall vocabulary knowledge. *System*, *87*, 102161. doi:10.1016/j.system.2019.102161

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, *29*, 471–488. doi:10.1177/0265532211433033

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary tests with linear models. *Shiken Research Bulletin*, *16*, 2–7.

Tang, E. (2007). An exploration study of the English vocabulary size of Hong Kong primary and junior secondary school students. *The Journal of Asia TEFL, 4*(1), 125–144.

Thornbury, S. (2002). *How to teach vocabulary*. Essex, UK: Pearson Longman.

University of Cambridge Local Examination Syndicate. (2001). *Quick Placement Test*. Oxford, UK: Oxford University Press.

Yue, Y., & Fan, S. (2016). Measurement of vocabulary knowledge: Problems and solutions. In S. Fan & J. Fieldings-Wells (Eds.), *What is next in educational research?* (pp. 171–182). Rotterdam, Netherlands: Sense Publishers.

Zhang, X. (2013). The I Don't Know option in the Vocabulary Size Test. *TESOL Quarterly*, *47*, 790–811. doi:10.1002/tesq.98