Effects of Test Format in Assessing L2 Vocabulary Knowledge and Skills

Effects of Test Format in Assessing L2 Vocabulary Knowledge and Skills

Siwon Park

Yasuko Ito

Megumi Sugita

Abstract

In this paper, we present preliminary results of a study in which we examined the relative contribution of English learners' vocabulary to predicting their reading and grammar knowledge, by employing two different formats of vocabulary tests that require active and passive recognition, respectively. We administered a series of English tests to over 820 university students, including TOEFL ITP, a reading test, a grammar test, and a vocabulary test with 80 items of two different formats. The target 80 vocabulary items were selected from Level 2 to Level 6 of JACET 8000. We analyzed the test data using statistical techniques in order to observe the relationships between the vocabulary levels and the language skills, and between the test format, the vocabulary level, and the language skills. We also examined the relative contribution of the vocabulary estimated in different item formats to predicting students' performance on other skills tests. The findings suggest that there was a very strong trait effect, dominating both methods when they are presented together with the vocabulary trait on the common factor structure. Also, the contributions of each method to predicting skills' performance were not consistent across the traits of grammar and reading.

Introduction

In the field of L2 education and, in particular, in L2 assessment, vocabulary has enjoyed popularity as a topic of research (Nation, 2006; Read, 2000, 2004). In theoretical conceptualization, researchers have proposed differing views on the nature of L2 vocabulary and have suggested different approaches in understanding its knowledge components: multicomponent vs. trait (Chapelle, 1998; Meara, 1996; Nation, 1990; Read & Chapelle, 2001). Consequently, such views have contributed to different ways of operationalization of L2 vocabulary trait in assessing learners' vocabulary knowledge alone or in relation to other L2 skills (Laufer & Goldstein, 2004; Zhang, 2012).

Concerning the strength of L2 vocabulary knowledge, prior studies have shown that there are two dimensions for categorizing vocabulary knowledge (Laufer & Goldstein, 2004; Meara, 1996; Nation, 2001). The first dimension is whether the knowledge is active or passive. Active knowledge is considered productive, often demonstrated in writing and speaking, while passive knowledge is receptive, demonstrated in reading and listening. The other dimension is whether it is recollection or recognition. It is also associated with the test format, because recall knowledge is elicited by asking learners to spell out the answer either in their L1 or L2, whereas recognition knowledge includes having learners choose one of the given choices in a multiple-choice format. These two dimensions produce four categories, as shown in Table 1, and Table 2 shows what each category would look like as test item.

Table 1

Classification of Vocabulary Knowledge

	Recall	**Recognition**
Active (retrieval of form)	Supply the L2 word	Select the L2 word
Passive (retrieval of meaning)	Supply the L1 word	Select the L1 word

Table 2

Examples of Test Items from Sasao (2008)

Test format		Test item
Recall	Active	抽象的な a
	Passive	abstract
Recognition	Active	抽象的な A. absolute B. abstract C. agricultural D. alleged E. わからない [Don't know]
	Passive	abstract A. 全くの B. 抽象的な C. その時代の D. 目に明らかな E. わからない [Don't know]

Studying the strength of L2 learners' vocabulary knowledge using the four categories, Laufer and Goldstein (2004) found that active recall is the strongest form of knowledge, while passive recognition is the weakest. In other words, active recall would be the most difficult and passive recognition the easiest format for the learners. Webb (2008), whose participants were 83 native-speakers of Japanese at a university in Japan, examined receptive (passive) and productive (active) vocabulary sizes, and the results revealed that the learners' receptive vocabulary size was larger than the productive one. He further argues, "[t]he findings indicate that receptive vocabulary size might give some indication of productive vocabulary size. Learners who have a larger receptive vocabulary are likely to know more of those words productively than learners who have a smaller receptive vocabulary" (p. 91). Sasao (2008) measured vocabulary size using the four test formats given above, which yielded significantly different scores. Similar to Webb (2008), Sasao found that the size of passive vocabulary was larger than that of active vocabulary.

Purpose of research

The purpose of this study was to explore the relative contribution of L2 vocabulary knowledge to predicting L2 learners' reading and grammar performances, assessed by the two different formats of vocabulary tests: passive and active recognition. More specifically, the study examined: (a) if the two formats were divergent when presented with the trait of L2 vocabulary knowledge on a common factor structure, and (b) if their contribution to predicting English learners' performance of grammar and reading skills was consistent across the traits that represented them.

Previous research has shown that there was a fixed hierarchy among the four test formats when measuring L2 learners' vocabulary knowledge. As suggested in the previous research, it is hypothesized in the present study that active recognition is stronger than passive recognition in terms of the strength of L2 learners' vocabulary knowledge.

Methods

Participants

Participants in this study were 788 freshmen, 545 female, and 243 male, all majoring in English at a university in Japan. Since the tests were administered immediately prior to or upon their entrance to university, most students were aged around 18. Their TOEFL ITP scores ranged from 370 to 557.

Test instruments

Every year, incoming students take different English tests prior to and upon their entrance to university. The tests include TOEFL-ITP, a vocabulary test, a grammar test, and a reading test. The vocabulary test, the grammar test, and the reading test are all developed internally. More details of each test are given below. The four test formats were also used in Mochizuki (2012),

in which L2 recall (active recall) was found to be most difficult, and L1 recognition (passive recognition) the easiest, supporting Laufer and Goldstein (2004).

The vocabulary test was developed based on JACET 8000 (Aizawa, Ishikawa, & Murata, 2005). Target words were selected using a randomizer for each level of the book, from Level 2 to Level 6. Levels 1, 7, and 8 were not included in the test, because those levels were either too easy or too difficult. The test format was both active and passive recognition, as described in Table 3. There were 80 items in total, with 40 items in each test format: 5 items from Level 2, 10 items from each of the levels from 3 to 5, and 5 items from Level 6. Table 3 provides examples of test items.

Table 3

Examples of Test Items from the Vocabulary Test

Active recognition	Passive recognition
1. 埋め合わせる、補償する a) conform b) contempt c) covenant d) compensate	2. Academic a) 礼儀正しい b) 正確な c) 抽象的な d) 大学の

The grammar test was developed based on a grammar textbook titled “Forest.” This is a popular grammar book among high school students in Japan. Since the participants in the study were all freshmen, the majority of whom have just graduated from high school, many of them were already familiar with the book. There were two sections on the test, usage and written expression, which were similar to TOEFL-ITP Grammar Section.

The reading test was developed in accordance with CEFR-J (Tono, 2013). The test included various types of reading texts, such as a story, a recipe, and an announcement (Fujimura & Sugita, 2015; Park & Ito, 2015). TOEFL-ITP, an institutional version of TOEFL,

was also used in data collection, which included three sections of English listening, grammar, and reading.

Results and Discussion

In analyzing the test data, IBM SPSS 21.0 (IBM Corp, 2012) and EQS 6.1 for Windows (Bentler, 2004) were employed. Table 4 shows the descriptive statistics and the reliability coefficients in the last column that concern the test instruments used in the study.

Table 4

Descriptive Statistics (N=788)

Test		*k*	*M*	*SD*	Min-Max	*R*
TOEFL ITP	Grammar	40	41.10	5.22	31-56	
	Listening	50	45.60	3.70	31-57	
	Reading	50	43.35	5.02	31-57	
Institutional English Test	Reading	33	20.14	3.86	7-30	.78
	Grammar	40	.50	.13	15-90	.83
Level-based English Vocabulary Test 1 (Japanese-to-English Format)	Level 2	5	.96	.09	0.4-1.0	
	Level 3	10	.77	.16	0.2-1.0	
	Level 4	10	.58	.16	0.1-1.0	
	Level 5	10	.59	.17	0.1-1.0	
	Level 6	5	.66	.23	0-1.0	
	Total	40	.69	4.43	.35-1.0	.73
Level-based English Vocabulary Test 2 (English-to-Japanese Format)	Level 2	5	.97	.08	0.4-1.0	
	Level 3	10	.89	.12	0.1-1.0	
	Level 4	10	.56	.17	0.1-1.0	
	Level 5	10	.56	.18	0.1-1.0	
	Level 6	5	.66	.22	0-1.0	
	Total	40	.70	4.2	.38-.95	.72

We analyzed the data using two structural equation models and examined their model fits and factor loadings: 1) A model of two methods and one trait (against one with one-method and one-trait) and 2) another model of a complete factor structure that includes the traits and the measures of English vocabulary, reading, and grammar. Figure 1 exemplifies the first model of two methods, active and passive, on the left-hand side, and one factor, vocabulary knowledge, on the right-hand side. Each method factor loads onto four measurement variables, vocabulary tests of differing levels, while the vocabulary trait factor loads onto all the measurement variables in concert. We identified the model in Figure 1, using the data collected from 788 examinees who sat the exams administered for the study.

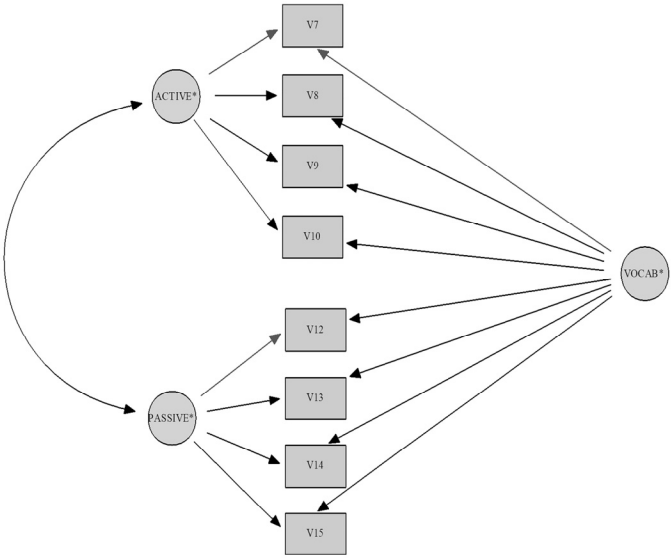

Figure 1 *SEM Model: Two-method and One-trait (N=788)*

Note. V=Vocabulary Test

The resulting model from the calibration is presented with model-data fit indices, relevant factor loadings, and a correlation coefficient between the two methods in Figure 2 below. The model fit indices of CFI=0.97 and RMSEA=0.02 demonstrate that the proposed factor structure is empirically supported by the data. There is a sufficient convergence between the model and the data, which leads to trustworthiness of the information provided in Figure 2.

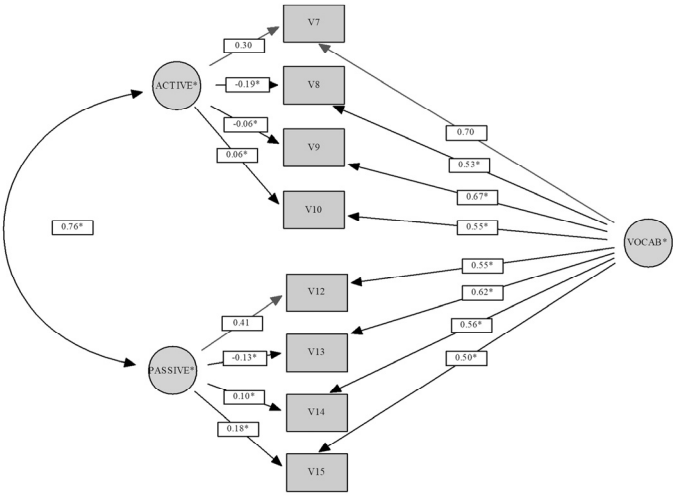

Figure 2 *SEM Model: Two-method and One-trait (N=788; Chi Sq.=33.22, p=0.02, CFI=0.97, RMSEA=0.02)*

Note. V=Vocabulary Test

While the correlation between the two method factors is relatively large at 0.76, the overall tendency concerning the factor loadings between the measurement variables and method and trait, respectively, indicates that there is a strong trait effect which outweighs that of the method. That is, the method loadings are mostly negligible, while those of trait to measurement variables

all fall in the range from 0.50 to 0.70. As the factor loadings of method to measurement variables are mostly negligible, the differential effects of the two methods in assessing L2 vocabulary knowledge cannot be examined. However, the relatively large size of the correlation coefficient 0.76 between the two method variables indicates a possibility of being convergent, i.e., the two methods agree in measuring the L2 vocabulary trait posited in this study (Byrne, 2006).

The second model that we examined in the study is presented in Figure 3, a model with a complete factor structure that includes the traits and the measures of English vocabulary, reading, and grammar knowledge. At the center of the model, four factors are presented: two methods of active and passive on the left, and two traits of grammar and reading on the right. Each method and trait is associated with the measurement variables. While the two methods are correlated, as the result of the first model suggested, the two traits are indirectly associated via the two methods.

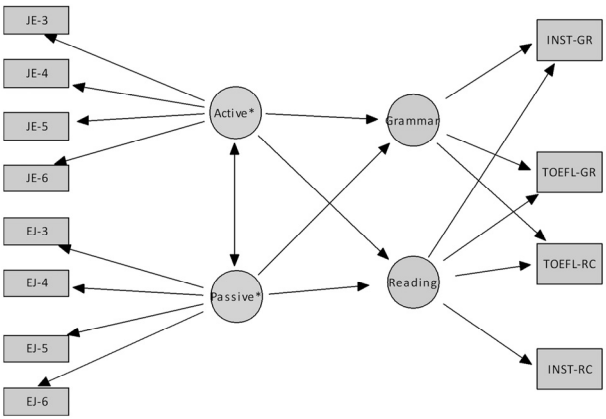

Figure 3

SEM Model of English Vocabulary and Skills Tests

Note. JE=Japanese-to-English Format Vocabulary Test; EJ = English-to-Japanese Format Vocabulary Test; INST-GR = Institutional Grammar Test

The model was estimated with the test data from 788 examinees. Figure 4 includes the information concerning the model fit, factor loadings from trait/method to measurement variables and from method to trait, and a correlation coefficient between active and passive methods.

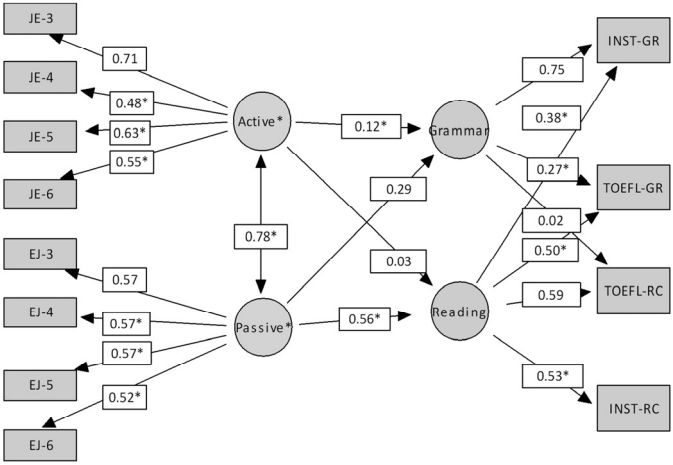

Figure 4

SEM Model of English Vocabulary and Skills Tests (N=788; Chi Sq.=119.58, p=0.00, CFI=0.97, RMSEA=0.03)

First, the correlation between the two methods is high and statistically significant at 0.78, demonstrating a convergent validity which shows that the two methods converge in assessing the L2 vocabulary traits operationalized by the eight measurement variables. This finding is consistent with the result from the first model.

After examining the association between the method and trait factors, the method operationalized by the passive test format demonstrated a stronger association with the reading trait (0.56) and the grammar trait (0.29, nonsig.), although the loading from passive to grammar appeared statistically non-significant. That is, passive recognition predicted grammar and reading to a greater degree than active recognition, and it demonstrated more predictive power to reading than to grammar. This finding is opposite to the ones from prior studies. One possible explanation of the finding may lie in the fact that the tests employed to assess L2 reading and grammar included items that promote passive recognition in paper and pencil format. If the test requires more active use of L2 vocabulary knowledge in test performance, the result may be different.

When both methods are presented together with the vocabulary trait on the common factor structure, they are identified equally well in relation to their measurement variables. However, the passive method (between 0.52 and 0.56) contributes to the measurement of the differing knowledge of L2 vocabulary level more consistently than the active method (between 0.48 and 0.71).

Finally, the contributions of trait factors, grammar, and reading to predicting their corresponding English skills were also inconsistent. While the reading trait demonstrated a stronger association with reading and grammar measurement variables (between 0.27 and 0.59), the grammar trait resulted only with the grammar measurement variables (0.75 and 0.27). The grammar trait showed a negligible association with the TOEFL-RC at 0.02. Although in its current model of the common factor structure we cannot conclusively identify the differential effects of grammar and reading traits onto the measurement variables, the grammar knowledge is, in fact, a part of the reading trait. That is, L2 reading performance may require the examinees to demonstrate the L2 reading, as well as L2 grammar knowledge.

Conclusions

The purpose of this study was to explore the relative contribution of L2 vocabulary knowledge to predicting L2 learners' reading and grammar performances, assessed by the two different formats of vocabulary tests: passive and active recognition. More specifically, the study examined if the two formats were divergent in assessing L2 vocabulary knowledge and if their contribution to predicting English learners' performance of grammar and reading skills was consistent across the traits in measurement.

The results from the first model indicated that the two formats, represented as different methods in the study, may tap comparable aspects of L2 vocabulary knowledge; as a method, they converge in assessing L2 vocabulary knowledge. However, the second model demonstrated that, unlike the common assumption from prior studies, passive recognition was more strongly associated with several L2 skills (e.g., grammar and reading) than active recognition. Therefore, in assessing L2 skills, a vocabulary test that employs passive recognition may be a better predictor of examinees' performance on reading and grammar than the one with active recognition. Also, the reading trait is more inclusive than the grammar trait, as the former appears to tap both L2 reading and grammar skills in measurement.

In this study, we did not identify a parsimonious model among the possible models through a model comparison approach. The primary purpose was not to identify the best fitting model, but rather to determine the common factor structure and the degree of associations among the factors and measurement variables. However, we recognize that the identification of a more parsimonious model and examinations of the factor structure and the associations of the factor relations may have provided more plausible arguments as to the research purpose of the differential effects of L2 vocabulary test methods in relation to L2 skills traits. That is one of the research avenues not yet reported in this preliminary study.

Reference

Bentler, P. M. (2004). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Byrne, B. M. (2006). *Structural equation modeling with EQS and EQS Windows: basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L.F. Bachman & A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge, England: Cambridge University Press.

Fujimura T., & Sugita, M. (2015). *Development of a level-specific test based on the CEFR-J reading scale*. Paper presented at the Japan Association of College English Teachers (JACET) Kanto Chapter 2015.

IBM Corp. Released 2012. *IBM SPSS Statistics for Windows, Version 21.0*. Armonk, NY: IBM Corp.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399-436.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-33). Cambridge: Cambridge university press.

Mochizuki, M. (2012). Receptive and productive knowledge of frequent and infrequent vocabulary. *Reitaku Review*, *18*, 64-79.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newsbury Hourse.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59-82.

Park, S., & Ito, Y. (2015). *Development of a CEFR-J based reading test*. Paper presented at Kanto-koshinetsu Association of Teachers of English (KATE) 2015 in Yamanashi.

Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.

Read J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146-161.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*, 1-32.

Sasao, Y. (2008). Estimating vocabulary size: Does test format make a difference? *JACET Journal*, *46*, 63-76.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79-95.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: a structural equation modeling study. *The Modern Language Journal, 96*, 558-575.

相澤一美、石川慎一郎、村田年（編集代表）(2005)．『「大学英語教育学会基本語リスト」に基づく JACET 8000 英単語』桐原書店

投野由紀夫（編）(2013)．『CAN-DO リスト作成・活用　英語到達度指標 CEFR-J ガイドブック』大修館書店

(2013)．『総合英語 Forest（7th edition）』桐原書店

(2013)．『総合英語 Forest（7th edition）完全準拠問題集　解いてトレーニング』桐原書店