English Language & Literature Teaching, Vol. 18, No. 2 Summer 2012

A Validity Study on the Vocabulary Grade Levels Test for Korean Elementary Students

Yousun Shin
(Pukyong National University)

Shin, Yousun. (2012). A validity study on the vocabulary grade levels test for Korean elementary students. *English Language & Literature Teaching, 18*(2), 125-147.

The primary goal of the study was to provide some preliminary validity evidence for the Vocabulary Grade Levels Test (Busan Metropolitan City Office of Education, 2009), which is designed to measure the receptive vocabulary knowledge of learners in L2. For the purpose of the current study, 327 participants at the elementary school participated in the study and were asked to take two different vocabulary tests. Namely, a Vocabulary Size Test (Nation, 2001) and a Vocabulary Grade Levels Test. The data were analyzed using correlation in order to discover the relationship between these two types of tests. Following this, the Rasch analysis was conducted to examine the reliability and validity of the measurement in question. The data analysis showed that both grade separation reliability and item separation reliability were high, indicating that the Vocabulary Grade Levels Test well discriminates learners with a wide range of proficiency levels. The findings of the study are discussed, along with further improvements in order to ascertain the validity of this particular vocabulary test.

[vocabulary knowledge/vocabulary level test/Rasch model]

I. INTRODUCTION

It is a commonly accepted belief that acquiring a large vocabulary gradually takes place over the years for both native speakers and foreign language learners (Belgar, 2010). In addition, the need for learners to strengthen their vocabulary is one of the most basic language skills, due to the fact that lexical knowledge plays such an essential role in reading and listening. Because of the importance put on vocabulary size as an indicator of ability to deal with L2 reading comprehension, a large number of studies have been conducted to identify the vocabulary size of different learner groups (i.e., Beglar, 2010;

Laufer & Nation, 1995, 1999; Meara & Jones, 1988; Qian, 2002; Xing & Fulcher, 2007). These various studies have used vocabulary test batteries, including the Vocabulary Levels Test (Nation, 2001) and Eurocentres Vocabulary Size Test (Meara & Buxton, 1987; Meara & Jones, 1990).

Regarding the relationship between reading ability and vocabulary knowledge, reading comprehension ability at the early educational level is strongly correlated with vocabulary levels, and the gaps between learners with a small vocabulary and learners with a large vocabulary tend to widen as their educational levels go up (Biemiller & Slomin, 2001). By the same token, the differences in vocabulary acquisition observed in school students could make a substantial contribution to later reading comprehension (Chall, Jacobs, & Baldwin, 1990).

However, it is hard to find an appropriate test to measure receptive vocabulary size found in foreign language learners, though such a test could serve several important roles in foreign language instruction and curriculum (Lee-Ellis, 2009). For example, the results of a vocabulary size test could be used to determine the current vocabulary size of individuals or language groups at a certain educational level, to make mastery decisions on lexical objectives in a program, or to better understand the progress of the learners' lexical development over the years before placing them into another level of any given language program. Thus, current foreign language programs, including those found in Korea, have called for a test which is capable of accurately measuring the vocabulary size of learners at each educational level.

Given the close relationship between the vocabulary command of EFL learners and their ability to understand certain passages while reading, it is justifiable to develop and implement a vocabulary test which effectively evaluates the vocabulary knowledge of EFL learners, and then to validate it for further use. In reality, despite an increasing interest in vocabulary learning in the EFL setting, an adequate measure to assess learners' vocabulary knowledge hasn't been thoroughly developed as of yet in Korea. This paper aims to describe the initial effort to provide validity evidence for one such test that is designed to measure the vocabulary size of language learners at the elementary school level.

The Vocabulary Grade Levels Test examined in this paper is intended to measure learners' receptive vocabulary size as well as some of the essential phrasal expressions which are frequently used in daily life. The test has been locally developed and disseminated by the Busan Metropolitan City Office of Education since 2009. Accordingly, its practicality and empirical validity evidence should be ensued in order for administrators, teachers, and even learners themselves to utilize the result of the test. Thus, the primary goal of the study is two-fold: (1) to determine to what degree the test is correlated to the Vocabulary Size Test (Nation, 2001) and (2) to investigate whether or not it is well-functioned in the current foreign educational context.

II. LITERATURE REVIEW

The importance of vocabulary knowledge has been a focus of study among L2 researchers. Researchers have found that vocabulary knowledge in L2 makes a major contribution to reading comprehension in L2 (e.g., Carrell, 1988; Choi, 2007; Kim & Ryoo, 2009; Koda, 1989; Laufer, 1992; Park, Lee, & Kang, 2005). Particularly, Kim and Ryoo (2009) investigated the relationship between Korean learners' vocabulary profiles and their reading and writing proficiency. 107 university students who participated in this study were asked to write an argumentative essay on two topics. They were then asked to complete a reading comprehension task. The results indicated that learners' ability of using academic vocabulary was essentially determined by their reading proficiency not by their writing proficiency. The results of the study implicated that there is a strong correlation between reading proficiency and vocabulary level in Korean academic contexts. Recently, Shin, Chon, and Kim (2011) conducted a study to assess the vocabulary size of Korean high school learners at three different English proficiency groups. The results indicated that receptive vocabulary knowledge was as large as 6,000 words, and found to be significantly larger than productive vocabulary knowledge. They reasoned that Korean learners invested more time and effort than the ESL learners to acquire a similar amount of passive vocabulary through deliberate learning.

Additionally, particular attention has been paid to test the adequacy of the receptive vocabulary size test including the Vocabulary Levels Test. Li and MacGregor (2010) investigated the effectiveness of the test with Chinese university learners in Hong Kong. The participants obtained high scores for high-frequency words but scored poorly for low-frequency words. They suggested the possibility that the learners' low scores might have resulted from a deficit of the test because low-frequency level words are not representative of Hong Kong English vocabulary. The study drew the conclusion that the test words should be representative of the vocabulary used in the learners' linguistic environments in order to obtain accurate and useful estimates of vocabulary.

Next, a few studies in relation to the validation study of a vocabulary test has been conducted up to now (i.e., Beglar, 2010; Lee-Ellis, 2009; Xing & Fulcher, 2006). Beglar (2010) investigated a study to provide validity evidence for the Vocabulary Size Test with 140-item form. Nineteen native speakers of English and 178 native speakers of Japanese participated in the study and the data were analyzed with using Rasch model. The findings indicated that the participants were measured with a high degree of precision on multiple versions of the test and the majority of the items showed good fit to the Rasch model. This validation study of the Vocabulary Size Test showed that the test greatly extends the range of measurement provided by other measures of written receptive vocabulary size. As another example of validation study on the Nation's Vocabulary Levels Test, Read (1988)

analyzed the results of 81 students who took the test during a three-month intensive course in English for academic purposes. They were asked to complete the test two times shortly after they began the course, and at the end of the course. The results showed the mean scores were consistently higher at the end of the course and the same general patterns of declining scores were found across the frequency levels (2,000, 3,000, 5,000, University Word List, 10,000 word level). He also found a substantial degree of implicational scaling across the five frequency levels. Put it another way, learners knew more of the items at the 2,000 word level than they did at the 3,000 word level as a general rule.

On the other hand, Xing and Fulcher (2007) examined the reliability of the two versions of Vocabulary Levels Test at the 5,000 word level. The data analysis of the study showed that Version A and B at the 5,000 word level were highly correlated and highly reliable. However, the item facility values of the Version B contain too many more difficult items to be parallel between the two versions. The researchers suggested that changes need to be made to the test before it is used in future vocabulary growth studies.

Unlike the above studies, Lee-Ellis (2009) developed and validated a 30-minute Korean C-Test[1] which was designed to assess Korean as a second language (KSL) learners' general language proficiency. 37 learners of Korean participated in this study and they were asked to complete the test and the self-assessment questionnaire. Rasch measurement statistics was used to analyze the data and examine the reliability and concurrent validity of the test. The developed test demonstrated high reliability and validity indices, which implies that the C-Test could be a reliable proficiency indicator.

In sum, despite the importance of developing a valid vocabulary test and examining the validity of it for further practical use, it is found that a few studies relevant to the issue has been done in either vocabulary tests or other general proficiency language tests. Thus, the present study attempts to investigate the reliability and concurrent validity of the developed vocabulary test to ensure its practicality and usefulness as an indicator for vocabulary growth of Korean learners.

III. METHODS

1. Participants

327 Korean-speaking elementary school students ranging from 3rd grade to 6th grade in two school districts participated in the study during their regular English class time

[1] The C-test was developed by Klein-Braley (1981), which was proposed as an alternative to the cloze test procedure.

(Female: 141, Male: 186). The participants were all enrolled in public schools at the time of the study. English proficiency is highly regarded in Korea and thus the educational policy there (here) strives to reflect this by creating an atmosphere that attempts to improve the learners' ability to communicate in English as early as possible. Therefore, under the current educational curriculum students begin learning English as a foreign language as early as the 3rd grade. According to the latest national curriculum revision in 2008, 3rd and 4th grade students have 2 hours of English instruction per week, while 5th and 6th grade learners receive 3 hours of English instruction per week in public schools. Meanwhile, the students at the private school have studied English for a minimum of 3 years since their first grade and have been receiving English lessons five hours per week all across the grade levels.

2. Instruments

1) Vocabulary Size Test

One of the Vocabulary Size Tests developed by Nation (2001) was used to measure learners' receptive vocabulary size at three frequency levels. This test took into consideration the learners' age along with the stage of their language development, and categorized them as follows: the 1,000 word level, the 2,000 word level, and the 3,000 word level. In the test, students were asked to match the vocabulary in the sentence with the meanings provided in Korean, as given in the following example:

Period: It was a difficult period.
a. 질문
b. 기간
c. 해야 할 일
d. 책

The test is intended to exhibit estimates of the vocabulary size at three different levels. Therefore, it is useful to diagnose vocabulary gaps among students. In the present study, the scores on this test were used to identify the students' vocabulary levels at each grade level. There were ten clusters at each level of the Vocabulary Size Test, and all of the items were the multiple choice types, each question consisting of four choices. For every correct answer, a student was awarded a single point, with the maximum score totaling 30 points.

2) Vocabulary Grade Levels Test (8th grade)

The Vocabulary Grade Levels Test was developed by the Busan Metropolitan City Office of Education (2009) to measure the learners' ability at each educational grade as

well, ranging from the 5th grade in elementary school to 3rd grade in high school. Each of these level tests was composed of 25 multiple-choice items with five possible options. Each question was worth four points, and the total score amounted to 100 points. The test was designed to measure how much students know about essential vocabulary at a certain grade level in percentage terms. The cutoff mark was 80% of the total score, which means that the students with 80% or above this score were awarded a certification at each grade level. For instance, 5th grade students start to take the 9th grade level vocabulary test, and 6th grade students are supposed to correspond to the 8th grade level test in terms of their general vocabulary knowledge. However, there has been not published any critical reviews or psychometric information on the test.

Unlike the Vocabulary Size Test developed by Nation (2001), this test includes different test formats such as matching a picture with its proper word, selecting a correct corresponding Korean word based on a given English word, selecting a word whose characteristics is not grouped with other options, or finding an appropriate English expression. All of the participants in this study took an 8th grade level test which was originally intended to measure 6th graders at elementary school. The reason for using the same test for all levels of the participants was that the test results from different grade levels would provide some pedagogical implications about the test, or even practical directions for improving the quality of the vocabulary test in question. Moreover, the results will help teachers and researchers to determine the "readiness" of each grade level on vocabulary learning to some extent.

Table 1 provides detailed information about the Vocabulary Grade Levels Test including test titles, corresponding graders with the test levels, the numbers of word token and word type[2] at each frequency level of the exemplary tests administrated in 2010. All of this was analyzed by the Lexical Frequency Profile (LFP) (Laufer & Nation, 1995). This tool categorizes the words in a text according to which frequency band level each word belongs to: first 1,000 most-frequent level, second 1,000 most-frequent level, 570 most-frequent 'academic' words which are included in the University Word List (UWL), and the not-in-the-list level, which basically does not belong to any frequency level.

TABLE 1

Test by Grade Levels and the Results of LFP

Grade	Test title	1,000 levels		2,000 levels		UWL		Not in the List		Total words	
		Token	Types	token	types	Token	Types	token	types	token	types
5th grade	9th grade	146	94	29	27	0	0	25	25	200	146
6th grade	8th grade	154	109	41	28	0	0	5	4	200	141

[2] Word type is defined as a base form (lemmatization), including all inflected and derived forms of a word family while word token is defined as any functional word and lexical word without lemmatization (Laufer, 1992).

7th grade	7th grade	260	174	62	61	7	7	50	47	379	289
8th grade	6th grade	296	195	68	68	13	13	30	30	407	306
9th grade	5th grade	254	163	77	75	35	33	31	31	397	302

In the case of 9th grade and 8th grade level tests, most of the words appeared in the test belonging to the 1,000 and 2,000 frequency levels. However, some of the not-in-the-list levels words are also shown in the two tests.

At the same time, however, I need to examine the proportion of frequency levels of the recommended word lists, which was stated and published by the Ministry of Education, Science and Technology (2008). This list has been frequently used in the past and is one of the most convenient vocabulary pools available when developing a language test that takes into consideration learners' educational level. The list for elementary school students consists of 736 words and its distribution of word types is as follows:

TABLE 2

The Results of LFP for the Recommended Word List

Level	1,000 types	2,000 types	UWL types	NIL types	Total words
Elementary school level	482 (65.4%)	173 (23.4%)	3 (.004%)	78 (10.8%)	736

The 1,000 and 2,000 frequency word levels are comprised of more than 88 % of the total words in the test. Despite the fact that the word list has been developed to increase elementary school students' vocabulary, the list included some of UWL and not-in-the-list words as well shown in the Table 2. Because the words in the LFP were selected based on frequency counts from dated word lists of American words, however, there is a high possibility that the sets of word list might not be representative of current English context in Korean (Li & MacGregor, 2010).

3. Procedures

All students participating in the study were given two measures over a two week period. First, the students completed a 30-item Vocabulary Size Test (see Appendix 1) to determine their vocabulary knowledge at three frequency levels. Next, the second Vocabulary Grade Levels Test (see Appendix 2) was administered during their regular English class time. The participants were given 20 minutes to complete each test.

IV. RESULTS AND DISCUSSIONS

Table 3 presents means, standards deviations, and score distribution statistics for the two vocabulary tests with the graphic presentation of the scores in Figure 1.

TABLE 3

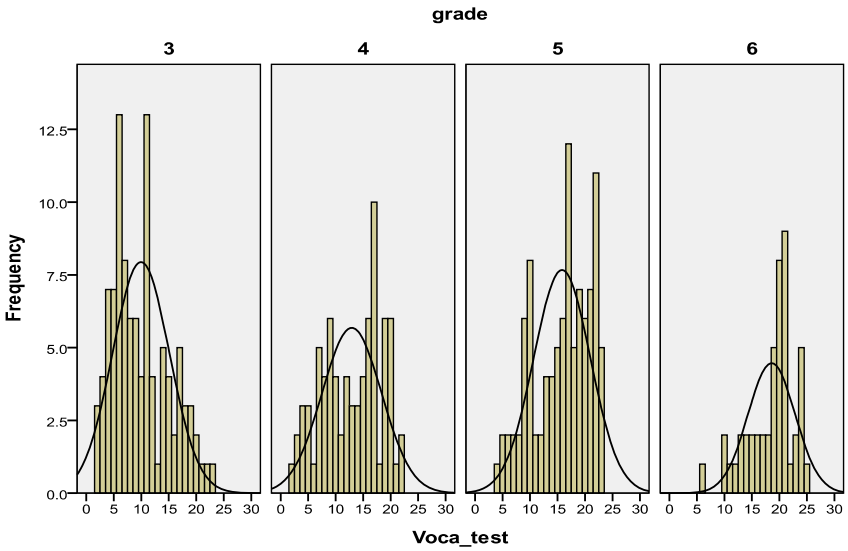
Descriptive Analyses of Two Vocabulary Tests (n=327)

Grade		N	Min	Max	M	SD	Skewness	
							Statistic	SE
3	NVST*	103	2	21	10.03	3.833	.743	.238
	VGLT**	103	2	23	9.95	5.172	.579	.238
4	NVST	77	4	22	11.48	3.905	.443	.274
	VGLT	77	2	22	12.94	5.408	-.237	.274
5	NVST	99	0	25	13.80	4.506	-.127	.243
	VGLT	99	4	23	15.82	5.152	-.470	.243
6	NVST	48	7	27	16.56	4.708	.036	.343
	VGLT	48	6	25	18.58	4.287	-.911	.343

*NVST: Nation's Vocabulary Size Test
**VGLT: Vocabulary Grade Levels Test

FIGURE 1

The Score Distribution of the Vocabulary Grade Levels Test



In spite of a certain degree of abnormalities, the score distribution of the Vocabulary Grade Levels Test indicated an opposite pattern between the 4th graders and the 6th graders. The patterns changed from the positively skewed distribution at the 4th grades to the negatively skewed one at the 5th and 6th grades. Considering that the Vocabulary Grade Levels Test is a criterion-referenced test, we can conclude from these data that the test is well-functioned at the 5th and 6th grade levels. In order to examine the differences among the grade levels and test items in more detail, I need to look at some of the other variables in this study via the Rasch analysis.

1. The Results of Correlations Between Two Vocabulary Tests

As a measure of the internal consistency of the tests, the Cronbach Alpha between the Vocabulary Size Test and the Vocabulary Grade Levels Test in this study was .865. As shown in Table 4, a significant relationship was found among the Vocabulary Size Test, the Vocabulary Grade Levels Test and the students' grade levels.

TABLE 4

Correlations among Two Vocabulary Tests

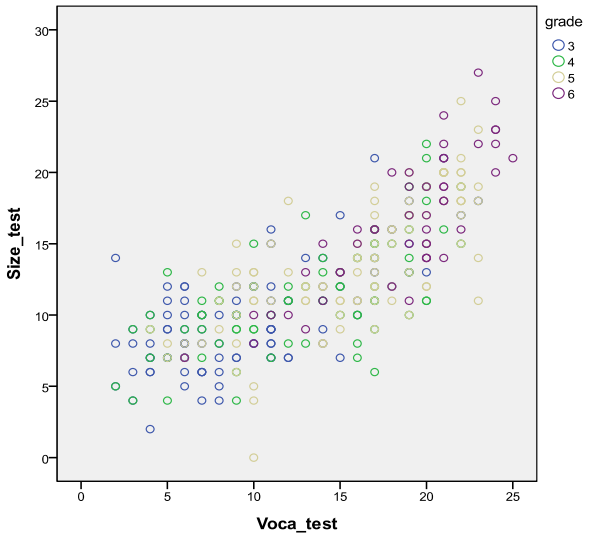
	1	2
1. Vocabulary Size Test	--	
2. Vocabulary Grade Level Test	.780**	--
3. Grade	.471**	.518**

**. Correlation is significant at the 0.01 level (2-tailed).

Regarding the relationship among the variables, all of the correlations summarized in Table 4 were found to be statistically significant. The Vocabulary Grade Levels Test was positively correlated with the Vocabulary Size Test and the students' grade level. That is, the higher the score the participants got in the Vocabulary Size Test, the higher the scores they received in the Vocabulary Grade Levels Test, and vice-versa.

FIGURE 2

The Scatterplot of Two Vocabulary Tests



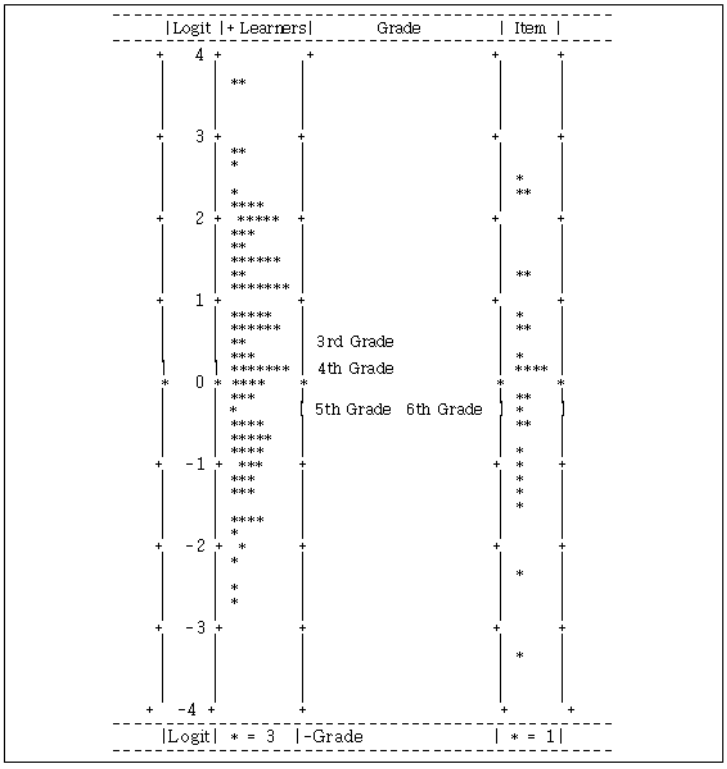
Similarly, the students in the higher grade levels tended to perform better either in the Vocabulary Size Test or the Vocabulary Grade Levels Test. Likewise, the positive correlations among the variables are clearly shown in the below scatter- plot shown in Figure 2.

2. The Results of the Rasch analysis

In order to better understand the interrelationships among these variables, the Rasch dichotomous model (Rasch, 1960) was employed. This particular model demonstrates that the likelihood of a particular score on an item from a particular learner in a particular grade can be predicted with mathematical certainty when three things are considered: the ability of the learner, the difficulty of the item and the proficiency level of the grade (McNamara, 1996).

FIGURE 3

Facet Map for the Vocabulary Grade Levels Test



Note: Each * in the first column represents approximately 3 persons. Each * in the fourth column represents approximately 1 item.

Figure 3 illustrates on average the relative abilities of the learners, the relative proficiency level of the grades, and the relative difficulty of the items in terms of the scale of probability developed in the analysis. There are four columns in Figure 3: one indicates the scale of measurement used, and the other three columns indicate each of the facets - *learner*, *grade* and *item*. The measurement scale in the first column called *logit* is used to report the estimates of the probabilities of the learners' responses while taking a test under the various conditions of measurement (ability, grade, item). It is readily apparent that the average item difficulty according to this scale is set to zero. Therefore, the items with negative signs are easier than average, while those with positive signs are more difficult than average. In the second column, there was a wide spread of ability and many different ability levels were identified. More than half of the students are situated above zero, which indicates that there are more able learners compared to those who fell below zero. Considering the proficiency level across all the graders in the third column, however, the results demonstrated that this test was simultaneously easier than average to both 5th grade and 6th grade students. As expected, the 3rd grade learners had more difficulty than the rest of the grade levels. Finally, in the fourth column, a wide range of item difficulty can be seen with respect to the difficulty of the items.

1) The Results by Grade Level

To take a closer look at the grade pattern distribution, the measurement report for grades is provided in Table 5. This pattern is ordered by grade level.

TABLE 5

Grade Measurement Report

Grade	Model		Infit	Outfit
	Measure(logit)	S.E.	MnSq	MnSq
3rd Grade	.55	.05	1.1	1.5
4th Grade	.08	.06	1.0	1.2
6th Grade	-.27	.08	.9	.7
5th Grade	-.37	.05	.9	.9
Mean(Count:4)	.00	.06	1.0	1.1
S.D.	.36	.01	.1	.3
RMSE (Model) .06 Adj S.D. .35 Separation 5.70 Reliability .97				
Fixed (all same) chi-square: 174.5 d.f.: 3 significance: .00				

Grade level is measured in *logit*, centered around zero, where positive signs indicate difficulty for the test and negative signs indicate easiness for the test. The report shows that 3rd grade students are slightly overwhelmed by the test while the test is too easy for 5th and

6th grade learners with negative numbers (-.27 and -.37, respectively). This left very little difference between the two grade levels. The infit and outfit mean square residuals provide information on how much consistency was demonstrated by each grade level students in the test. (Johnson & Lim, 2009; McNamara, 1996). Fit statistics is used to show the degree of parity between a pattern of responses observed for a grade level for each particular item and the modeled expectation (Lee-Ellis, 2009). Unfortunately, there is no clear cutoff depicting which values are too high or which are too low. According to McNamara (1996), values within two standard deviations from the mean should be the benchmark, as it were, which, in this case, would mean that value ranging from 0.8 to 1.2 would qualify as an infit statistic, while those from 0.5 to 1.7 would qualify as an outfit statistic. All of the grade levels fall safely within both standards set by McNamara. Overall, the fit statistics implies that all grade level learners showed reasonable consistency in the test.

The reliability statistic provided by the FACETS analysis indicates the degree to which the analysis reliably discriminates among different levels of grades (McNamara, 1996; Weigle, 1998). The reliability index is .97 for all the grade levels, suggesting that the analysis is reliably and accurately partitioning grade levels into different proficiency groups. Lastly, the chi-square of 174.5 with 3 d.f. is significant at $p<.00$, indicating that the proficiency of the grade levels are not equally the same.

2) The Results by Item Difficulty

To observe the quality of the items of the test, item fit statistics were examined. Table 6 summarizes the FACETS results for item difficulty. More than half of the items clustered around the mean (0 logit), suggesting that only a few items were particularly difficult (i.e., item 24, 21, 20, 22) or easy (i.e. item 5, 3). More specifically, item 3 (-3.28) was the only item whose easiness was more than two standard deviations from the mean. Items whose infit statistics are higher than two standard deviations from the mean indicate some unpredictability based on the modeled expectation. On the other hand, items with over two standard deviations of outfit values show the lack of variability among the items. That is, the items weren't successful in differentiating the learners with their different proficiency levels.

TABLE 6

Item Measurement Report

Item No.	Model		Infit	Outfit
	Measure (logit)	S.E.	MnSq	MnSq
24	2.53	.17	1.1	2.8
21	2.39	.16	1.5	4.5
20	2.27	.16	1.0	1.3

22	1.33	.14	1.3	1.5
25	1.33	.14	1.2	1.4
9	.91	.14	1.5	1.8
18	.62	.13	1.3	1.4
23	.62	.13	1.1	1.2
11	.36	.13	.7	.7
14	.24	.13	.8	.7
12	.20	.13	.8	.8
15	.13	.13	1.1	1.1
16	.13	.13	.9	.8
17	-.09	.14	.8	.7
8	-.11	.14	1.1	1.0
10	-.28	.14	.7	.6
6	-.51	.14	.8	.7
7	-.52	.14	.9	.8
4	-.86	.14	.8	.6
19	-.99	.15	.8	.6
1	-1.25	.15	.8	.7
2	-1.29	.15	1.2	1.6
13	-1.53	.16	.9	.7
5	-2.33	.18	.8	.5
3	-3.28	.25	.8	.3
Mean(Count:25)	.00	.15	1.0	1.2
S.D.	1.37	.02	.2	.9
RMSE (Model) .15 Adj S.D. 1.36 Separation 9.04 Reliability .99				
Fixed (all same) chi-square: 1608.8 d.f.: 24 significance: .00				

For the item difficulty, acceptable infit value ranges were from 0.6 to 1.4, while outfit value ranges were from -0.6 to 3.0. As can be seen, all of the items fall safely within both of the set standards except items 21 and items 9. These two items failed to discriminate the learner abilities in a manner consistent with other items. Thus, it appears that items 20 and items 18 need to be revised in order to more effectively distinguish learner abilities.

The reliability is .99 for item difficulty, indicating that the analysis is considerably reliable throughout the items. The chi-square of 1,608.8 with 24 d.f. is significant at $p<.00$, suggesting that all items are not equally difficult. As indicated by the grade separation reliability of .97 and item separation reliability of .99, the test items seem to function well in discriminating learners by their abilities.

V. CONCLUSIONS

The purpose of the present study was to provide some validity evidence for a test developed by the local educational administration in Korea by means of the correlational study and the Rasch analysis. A summary of the answers for the current research questions reads as follows: First, the results have shown that the Vocabulary Size Test and The Vocabulary Grade Levels Test are highly inter-correlated. The strong correlation between

the two tests implies that the two tests seem to measure the same traits associated with the vocabulary knowledge of learners. Additionally, the results of the Rasch analysis indicate that both grade separation reliability and item separation reliability are high. Furthermore, the Rasch analysis provided useful information that will go a long way in improving the quality of this test. In particular, the manner in which "misfitting" items can now be more easily identified from the analysis and therefore omitted from the test altogether, or revised for further use.

In summary, the Vocabulary Grade Levels Test provides a quick, reliable, and inexpensive way to measure a young learners' vocabulary knowledge. It is relatively easy to develop, provided that a test developer first carefully selects the vocabulary used in each grade level. Furthermore, administering this sort of test is rather quick and effortless, as is the scoring. Taking into consideration the high practicality of this test, the Vocabulary Grade Levels Test can be an effective tool to measure students' vocabulary knowledge, potentially providing reliable and accurate estimates of students' proficiency as seen at different educational levels.

However, it is obvious that the test needs further improvement to ascertain not only its reliability, but also its validity. It is noticeable that the Vocabulary Grade Levels Test demonstrates a wide variety of item formats compared to the unified test format of the Vocabulary Size Test (Nation, 2001). This fact might serve to have a harmful impact on the reliability of the measure. Bachman (1990) stated that by standardizing the input and response types, any given language test will yield highly reliable results. For instance, this could mean using a single testing format, say, a cloze test or a multiple choice-item test, requiring the exact same type of response from all the test takers, or even with vocabulary of the same frequency level. The complexity of the test format identified in the Vocabulary Grade Levels Test might lead to issues concerning the specific abilities that an educator is intent on measuring. It is highly conceivable, for instance, that this sort of test unintentionally measures learners' grammatical knowledge or use of idiomatic expressions, neither of which appears to be relevant to his or her vocabulary knowledge. Furthermore, the important factors of test formats and test methods also affect test interpretation, so that an attempt to standardize the test methods is required in order to increase the reliability of the test and subsequently the validity of the test (Bachman, 1990; Bachman & Palmer, 1996).

Another limitation of the test is that the vocabulary used in each test level is selected based on the recommended word list published by the Ministry of Education, Science and Technology (2008). In order to obtain accurate and useful estimates of receptive vocabulary knowledge, tests need to firmly establish first that the test words are indeed representative of the vocabulary used in the linguistic and learning environments relevant to the learners (Li & MacGregor, 2010). Unfortunately, the paucity of research to-date

concerning receptive vocabulary knowledge in a foreign language context, along with the general lack of knowledge about test design, is impeding our progress in this area. As a result, test specifications, and even a vocabulary selection process needs to be clearly articulated and open to the public.

Finally, a great deal of caution is required when interpreting these results, particularly in regards to the high separation reliability issue which was based on the Rasch analysis. Each grade level test is supposed to have a corresponding grade level of students (i.e., 9th grade level test – 5th graders at elementary level, 8th grade level test – 6th graders). In this present study, however, the proficiency range of the participants was widely spread from 3rd graders to 6th graders when taking an 8th grade level test. And so this and this alone probably led to the high item separation reliability. If the same test were administrated again, but this time to a specific learner group, for example 6th graders, the reliability would be lower (Lee-Ellis, 2009).

In closing, this study provides evidence that the Vocabulary Grade Levels Test can be safely used as an effective tool to measure the vocabulary knowledge of EFL learners. However, the test is certainly in need of some revision, especially at the item level. Nevertheless, a little more research into the selection process of the vocabulary that is to be used in such a test, as well as more research into the effect of the test formats, will undoubtedly shed some much needed light on the validity of this test. Consequently, we may become better equipped as educators to measure the language ability of EFL learners, here in Korea and in other countries around the world.

REFERENCES

Bachman, L. F. (1990). *Fundamental consideration in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101-118.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*, 498-520.

Carrell, P. L. (1988). SLA and classroom instruction: Reading. *Annual Review of Applied Linguistics*, *9*, 223-242.

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind.* Cambridge. MA: Harvard University Press.

Choi, S-M. (2007). How derivational prefix instruction impacts incidental vocabulary acquisition and reading comprehension. *English Language & Literature Teaching, 13*(3), 1-22.

Kim, S-Y., & Ryoo, Y-S. (2009). Korean college students' vocabulary profiles as predictors of English reading and writing proficiency. *Multimedia-Assisted Language Learning, 12*(3), 93-115.

Klein-Braley, C. (1981). *Empirical investigation of cloze tests*. Unpublished doctoral dissertation, University of Duisburg, Duisburg.

Koda, K. (1989). The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Language Annals, 22*(6), 529-540.

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: MacMillan.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307-322.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*, 33-51.

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

Li, L., & MacGregor, L. J. (2011). Investigating the receptive vocabulary size of university-level Chinese learners of English: How suitable is the Vocabulary Levels Test? *Language and Education, 24*(3), 239-249.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 142-154.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell. (Ed.), *Applied linguistics in society* (pp. 80-87). London: Centre for Information on Language Teaching and Research.

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

Park, J-W., Lee, G-I., & Kang, M-S. (2005). The effects of cognitive style and vocabulary learning strategies on students' achievements in web-based learning. *English Language & Literature Teaching, 11*(4), 21-47.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: The Danish Institute of Educational Research.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC*

Journal, *19*, 12-25.

Shin, D., Chon, Y. V., & Kim, H. (2011). Receptive and productive vocabulary sizes of high school learners: What next for the basicword list? *English Teaching*, *66*(3), 123-148.

The Ministry of Education, Science, and Technology (2008). Korean national school curriculum – English. Seoul: Author.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of vocabulary levels tests. *System*, *35*, 182-191.

APPENDIX 1

Vocabulary Size Test (Korean version)

학년: _____ 이름: _____

First 1000

1. see: They **saw** it.
   a. 잘랐다 b. 기다렸다 c. 보았다 d. 시작했다

2. time: They have a lot of **time**.
   a. 돈 b. 음식 c. 시간 d. 친구들

3. period: It was a difficult **period**.
   a. 질문 b. 기간 c. 해야 할 일 d. 책

4. figure: Is this the right **figure**?
   a. 대답 b. 장소 c. 시간 d. 숫자

5. poor: We are **poor**.
   a. 돈이 없다 b. 행복하다
   c. 매우 흥미 있다 d. 힘들게 일하는 것을 싫어한다

6. drive: He **drives** fast.
   a. 수영하다 b. 배우다 c. 공을 던지다 d. 차를 운전하다

7. jump: She tried to **jump**.
   a. 물 위에 누워있다 b. 갑자기 뛰어오르다
   c. 길가에 차를 세우다 d. 아주 빨리 움직이다

8. shoe: Where is my **shoe**?
a. 돌보는 사람 b. 금고, 저금통 c. 펜, 연필 d. 신발

9. standard: Her **standards** are very high.
a. 구두 뒷굽 b. 학교성적 c. 요구한 금액 d. 수준

10. basis: I don't understand the **basis**.
a. 이유 b. 단어들 c. 도로 표지판 d. 기본원리

Second 1000

1. maintain: Can they **maintain** it?
a. 유지하다 b. 확대시키다 c. 더 나은 것을 믿다 d. 얻다

2. stone: He sat on a **stone**.
a. 돌 b. 의자 c. 양탄자, 깔개 d. 나무

3. upset: I am **upset**.
a. 피곤한 b. 유명한 c. 부유한 d. 불행한

4. drawer: The **drawer** was empty.
a. 서랍 b. 차고 c. 냉장고 d. 동물의 집

5. patience: He has no **patience**.
a. 인내심 b. 여유시간 c. 믿음 d. 정직함

6. nil: His mark for that question was **nil**.
a. 매우 나쁜 b. 영점인 c. 매우 좋은 d. 중간인

7. pub: They went to the **pub**.
a. 술집 b. 도박장 c. 쇼핑센터 d. 수영장

8. circle: Make a **circle**.
a. 스케치 b. 빈 공간 c. 원(형) d. 큰 구멍

9. microphone: Please use the **microphone**.
a. 전자레인지 b. 마이크 c. 현미경 d. 휴대폰

10. pro: He's a **pro**.
a. 사립탐정 b. 미련한 사람 c. 기자 d. 직업으로 하는 운동선수

Third 1000

1. soldier: He is a **soldier.**
   a. 사업가 b. 학생 c. 금속 공예사 d. 군인

2. restore: It has been **restored**.
   a. 다시 말해졌다 b. 다른 사람에게 주어졌다
   c. 더 낮은 가격이 주어졌다 d. 다시 새것처럼 만들어졌다

3. jug: He was holding a **jug**.
   a. 주전자 b. 격식없는 대화 c. (챙이달린)모자 d. 폭탄

4. scrub: He is **scrubbing** it.
   a. 가는 선을 새기고 있다 b. 고치고 있다
   c. 깨끗하게 하기 위해 세게 문지르고 있다 d. 간단한 그림을 그리고 있다

5. dinosaur: The children were pretending to be **dinosaurs**.
   a. 해적 b. 요정 c. 용 d. 공룡

6. strap: He broke the **strap**.
   a. 약속 b. 뚜껑 c. 접시 d. 끈

7. pave: It was **paved**.
   a. 막혔다 b. 나눠져 있었다
   c. 금테가 둘러있었다 d. (도로가) 포장되어 있었다

8. dash: They **dashed** over it.
   a. 돌진하였다 b. 천천히 움직였다
   c. 싸웠다 d. 빨리 보았다

9. rove: He couldn't stop **roving**.
   a. 술 마시는 것 b. 배회하는 것
   c. 허밍 하는 것 d. 열심히 일하는 것

10. lonesome: He felt **lonesome**.
    a. 은혜를 모르는 b. 피곤한 c. 외로운 d. 힘이 넘치는

APPENDIX 2

필수영어단어급수인증제 8-I급 평가문제

제 학년 반 번 성명 _____

*** 점수는 문항당 4점입니다.**

[1~5] 다음 그림에 알맞은 단어를 고르시오

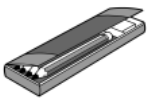
1.

① toe ② tooth ③ stomach
④ knee ⑤ chest

2.

① dentist ② teacher ③ nurse
④ police officer ⑤ fire fighter

3.



① ruler ② pencil case ③ eraser
④ bag ⑤ glue

4.

① sandwich ② pepper ③ kitchen
④ pork ⑤ vegetable

5.

① baseball ② bicycle ③ volleyball
④ soccer ⑤ tennis

[6~9] 다음 단어의 우리말 뜻으로 가장 적절한 것을 고르시오.

6. point : _____
① 도착하다 ② 끝내다 ③ 지적하다 ④ 배우다 ⑤ 공부하다

7. understand : _____
① 이해하다 ② 생각하다 ③ 쉬다 ④ 시도하다 ⑤ 사용하다

8. praise : _____
① 전화하다 ② 마시다 ③ 요리하다 ④ 자르다 ⑤ 칭찬하다

9. niece : _____
① 조카 ② 삼촌 ③ 이모/고모 ④ 조카딸 ⑤ 부모님

[10~12] 다음 중 단어의 성격이 **다른** 하나를 고르시오.

10. ① duck ② elephant ③ goat ④ back ⑤ dolphin

11. ① son ② mother ③ thumb ④ grandfather ⑤ daughter

12. ① socks ② pants ③ cap ④ skirt ⑤ shoulder

[13~16] 다음 문장의 빈 칸에 가장 적절한 단어를 고르시오.

13. I like this straw _____ . (나는 이 밀짚모자가 마음에 들어.)
① gloves ② hat ③ boots ④ jacket ⑤ shoes

14. Can I _____ why? (왜 그런지 물어봐도 될까요?)
① answer ② teach ③ ask ④ read ⑤ talk

15. _____ the steak, please. (스테이크 좀 잘라주세요.)
① Drink ② Fill ③ Cook ④ Eat ⑤ Chop

16. I _____ your help. (전 당신의 도움이 필요해요.)
① need ② count ③ teach ④ use ⑤ change

17. Where do I _____? (어디에서 돈을 지불합니까?)
① ring ② pay ③ try ④ understand ⑤ finish

[18~22] 다음 문장의 빈 칸에 가장 적절한 단어를 고르시오.

18. On my birthdays, I _____ some presents.

① give ② get ③ borrow ④ sell ⑤ rest

19. What time do you have _____?

① lunch ② elbow ③ pencil ④ scissors ⑤ doctor

20.

He is carrying a _____ of chicken. He is very busy but looks happy.

① fork ② knife ③ breakfast ④ plate ⑤ dinner

21.

A: Would you like to eat some _____ cutlets? B : Yes, I'd love to. Thank you.

① hot dog ② fork ③ pork ④ pepper ⑤ juice

22. I am so thirsty.

I will ____ a glass with milk and I will _____ it.

① cook/ fill ② chop/ cook ③ fill/ drink ④ drink/ eat ⑤ eat/ cook

23. 다음 그림에 맞는 문장을 완성할 때 알맞은 단어로 짝지어진 것을 고르시오.

I want to change my _____ to _____.

① pay/ coins ② a bill/ count ③ need/ a bill ④ count/ try ⑤ coins/ a bill

24. 다음 중 관계가 나머지와 다른 것을 고르시오.

① speak - talk ② ask - answer ③ teach – learn ④ give - get ⑤ take off - wear

25. 다음 중 의미상 어색한 문장을 고르시오.

① Take off your wet pencil.
② I got my pants dirty.
③ I like ducks, but I don't like snakes.
④ My father has big shoulders.
⑤ I want a pair of socks.

Examples in: English
Applicable Languages: English
Applicable Levels: Primary/Secondary

Yousun Shin
Division of English Language & Literature
Pukyong National University
599-1 Daeyon-dong, Nam-gu
Busan, 608-737 Korea
CP: 010-4760-0914
Email: yousun-shin@pknu.ac.kr

Received in March, 2012
Reviewed in May, 2012
Revised version received in June, 2012