Analysing the Reliability of Vocabulary Practice Tests

Kayoko Kinshi

1. Introduction

Vocabulary knowledge plays an important role in language learning. In order to discover the meaning and usage of the word, it is useful to use English paper/electronic dictionaries. However, the second language learners in Japan tend to almost entirely use bilingual English dictionaries in their study. For one thing, it is easy to have a quick look at the meaning of the corresponding word in Japanese; for another, even if they try to use monolingual dictionary to look up the unknown word, they often find it difficult to understand English definitions of it. This is because of the lack of their knowledge of basic defining words, which are essential in order to use monolingual English dictionaries. According to Nation (2008: 114), in order for learners to use monolingual dictionary, they need to have a vocabulary of at least 2,000 words. This is because of the fact that the number of defining vocabulary items used in the monolingual dictionary is around 2,000. Therefore, the Japanese university need to learn the defining vocabulary items to fill the gap, and this in turn necessitates the development of vocabulary practice test for learning defining vocabulary.

This paper discusses the reliability of the vocabulary practice tests which are developed for university students in Japan. The word items in these tests are exclusively designed for learning the defining vocabulary used in monolingual English dictionaries. The notion of reliability is referred to as “the consistency of the scores obtained” (Fraenkel, J. R., & Wallen, N. E. 1993: 146). If a test keeps giving the same result for the same participant over time, it can be said to serve the purpose of reliability. In this paper, the following two points are mainly discussed: the design of the revised vocabulary tests that would make the tests reliable, and the statistical results of test scores gained from using SPSS.

Section 2 overviews the background of the vocabulary practice tests, and some of their statistical results. Section 3 begins by addressing the process of revising the vocabulary test, and presents the test procedures. Section 4 analyses the findings, mainly focusing on the results from statistic data. Section 5 discusses the features of the revised vocabulary tests and statistical results in terms of reliability. Finally, the conclusion gives a brief summary and suggestions for further study.

2. Preceding Study

The design of the vocabulary tests discussed in this paper begins with an idea of compiling a list of vocabulary item for Japanese university students to learn (Kinshi 2009; Okada et al. 2009). The list was made with the defining vocabulary listed in two of the three major English monolingual dictionaries: *the Oxford Advanced Learner's Dictionary* (2000), *the Cambridge International Dictionary of English* (1995), and *the Longman Dictionary of Contemporary English* (2000). Out of a total of 384 headwords, such non-content words as numerals, pronouns, and auxiliaries were deleted, resulting in 300 content words. The 300 word items are then developed into two types of fill-in-the-blank tests, a cloze type and a paraphrase type, each of which has 5 tests with 60 words respectively (totaling 10 tests). Each sheet of the two series of five vocabulary practice tests ("VP 2008" hereafter) is designed to test 60 words, half of which are devised to be chosen as answers, and the other half are distracters. In the classroom, a list of 60 words is distributed two weeks before the VP 2008 is conducted. As Matsui et al. (2004: 101) reports, one of the effective way for learners to study vocabulary is to present "a clearly defined word list . . . along with concrete steps toward achieving the study target." Moreover, Nation (2001: 74) mentions that repetition is important to vocabulary learning in that it "adds to the quality of knowledge and also to the quantity or strength of this knowledge." Therefore, if learners take two types of the VP 2008 with the same 300 words, they are bound to expand their vocabulary knowledge.

Examining the results of the VP 2008 conducted on 178 Japanese students, Okada et al. (2009) reveals that there is an overall score improvement in the second series (i.e., a paraphrase type) over the first series (i.e., a cloze type). The mean score of each series rises from 25.46 to 26.75, out of 30. The Pearson correlation coefficient between the mean scores of the two series is .81 ($p<.001$), which is highly correlated. Moreover, they conduct the C-Tests (the Pre-and Post-Test, respectively), which are developed by the Writing Research Group of the JACET Kansai Chapter (1995, 1998). The C-Test consists of 4 paragraphs of 48-65 words, in which the last half of every other word is missing as shown in “Take off your shoes when you enter a Japanese house.” The students need to fill in the underlined part, which is devised to answer the half, or half plus one letter of each test word. The students have to identify the word form (singular-plural, present-past, etc.) as well as the meaning to fill in the blank. Each C-Test is scored out of 100 (two points for 50 items each). For such responses that are correct in meaning but contain a structural error, half of a mark (one point) is subtracted from the score.

As the result of the tests, the correlation coefficients between the two series of mean scores of the VP 2008 and the two C-Tests (the Pre-and Post-Test) are .59 and .68 ($p<.001$) respectively, which represent moderately high correlation figures. These results would be more highly correlated if the test format of the VP 2008 and the C-Test were designed to assess the similar vocabulary ability. That is, whilst the former is categorized as a “relatively decontextualized” test where context is “largely eliminated,” the latter measures “controlled productive ability” in which “there is a clear need to make use of contextual clues” (Okada et al. 2009: 12). Therefore, in order to assess the aspects of such word knowledge as the inflectional or derivational forms of words, with the two types of the tests, i.e., the vocabulary test and the C-Test, the VP 2008 should be revised to make it a more context-dependent test.

This study is designed to address the following two research questions:

⑴ Is there any improvement in students’ scores of the C-Tests?

⑵ Are there any significant relationships between the revised vocabulary tests and the C-Tests?

３．Method

3. 1 The Design of the Revised Vocabulary Practice Test

In response to the results of the study in the previous section, the VP 2010 was developed as a revised version. The purpose of the tests is to design more context-dependent tests, in which the learners have to fill in the incomplete words using the context clues and initial letters. The new version, the VP 2010, differs from the previous one in the following points: (1) the number of words in each list is reduced from 60 to 30; (2) each test is composed of 15 cloze type and 15 paraphrase type items, resulting in 10 tests with 30 words each; (3) 74 out of 300 words (about 25%) in the lists are changed into other forms on the word lists, for example *absent* instead of *absence*, or *benefit* for *beneficial*.

The significant difference between the VP 2010 and the VP 2008 is that the revised version is not a multiple choice type, but a fill-in-the-blank one, supplying the missing latter words. In addition, the initial letters are listed at random for every five questions. When taking the test, the learners would need to pay careful attention to the sentence context and not just use the initial letters alone as the clue for finding the correct words. A sample test from the VP 2010 is shown in Figure 1 below.

［The Cloze Type］
con ________ 1. If you're feeling cold here in the garden, we can go [   ].
cra ________ 2. If you think that today's topic does not [   ] you, you do not have to stay.
indo _______ 3. Tom is not a good student. He is too [   ] to do his homework.
la _________ 4. I heard a loud [   ] outside. I thought there was a car accident.
wed _______ 5. I have been invited to my friends' [   ] next week.
［The Paraphrase Type］
bact _______ 1. not knowing where you are or what is happening around you
ble ________ 2. very small living things that sometimes cause disease
ch _________ 3. the bottom part of your face, below your mouth
cont _______ 4. an official writing agreement between two or more people
uncon _____ 5. to lose blood

Figure 1　Sample Tests from VP 2010

3. 2　Test Procedure

The participants are 45 first- and second-year university students in Japan. They are all English majors and enroll in the English grammar class. While doing some grammar exercises in the textbook and writing a short paragraph using the topic they learned, they are working for the vocabulary practice tests during the semester. Vocabulary practice tests (VP 2010) are conducted every week during the semester; that is, one test sheet with 30 words each, for 10 weeks. In order for students to prepare for the tests in advance, they were given an alphabetically-ordered list of 30 words each week. In addition, a sample test sheet is presented, letting them know that, for preparation, it is necessary to consult the monolingual English dictionaries and to check the definition, usage, and word family of the words on the list. Vocabulary tests are conducted two weeks after they receive the word list. The students are allowed to work for 10-15 minutes on each test, which is scored out of 30. At the completion of the test, they score each other's responses, allowing them to immediately check their errors. Then the instructor collects the test sheets in order to check the word items that receive the lower scores. Such items are presented to the students the following week, with some explanations by the instructor.

In order to measure the depth of the students' vocabulary knowledge, the C-tests, which are the same ones in the previous section, are conducted at the beginning and at the end of the semester. The time limit for each C-Test is 15 minutes. The tests collected and scored by the instructor are returned to the students to check the errors on both occasions. At the end of the term (the week 15), the personal files including the 10 test scores, the two C-Tests scores, and mean scores of each test, are delivered to each student to monitor their improvement.

4. Results

First, Table 1 shows the mean scores and standard deviation for each of the 10 tests of the VP 2010. Comparing each of the 10 tests, the result of Test 3 seems different; the mean score is 18.6 out of 30 points, which is relatively low, and its standard deviation is 7.17, which indicates that the distribution of Test 3 is widely spread out from the mean, compared to the other 9 tests. However, Cronbach's alpha coefficient among the mean scores of the 10 series is quite high (α = .97). Therefore, it can be said that there is high internal consistency among the 10 tests.

Table 1 Scores and Standard Deviation of VP 2010 (10 tests)

	N	Minimum	Maximum	Mean	Std. Deviation
Test 1	34	8	28	20.4	5.37
Test 2	39	9	30	22.0	5.82
Test 3	36	2	29	18.6	7.17
Test 4	35	4	30	22.7	5.96
Test 5	35	6	29	23.0	5.57
Test 6	32	6	28	20.1	5.74
Test 7	37	10	30	22.9	4.79
Test 8	40	9	30	24.4	5.56
Test 9	40	7	30	24.9	5.08
Test 10	39	10	30	23.2	4.82

Next, Table 2 indicates the mean scores and standard deviation for the Pre-and Post-C-Tests. It is apparent that the mean score of the Post-C-Test increases by 8.4 (=75.1-66.7). The standard deviation of the Post-C-Test (14.59) is smaller than that of the Pre-C-Test (18.26), which shows the distribution of the Post-C-Test scores is more concentrated around the mean score. In addition, the result of the Pearson correlation coefficient between the scores of two C-Tests is .67 ($p< .001$), and Cronbach's alpha coefficient results in an adequate level (α = .79).

Table 2　Scores and SD of Pre-and Post-C-Tests

	N	Minimum	Maximum	Mean	Std. Deviation
Pre-C-Test	42	20	96	66.7	18.26
Post-C-Test	42	34	96	75.1	14.59

As shown in Figure 2, boxplots represent the distribution of the Pre-and Post- C-Test scores. From the graph, the median of the Post-C-Test is higher, and the bottom whisker is shorter than that of the Pre-C-Test. It can be said that some students with lower scores in the Pre-C-Test made progress even though there still exist some outliers for each C-Test.

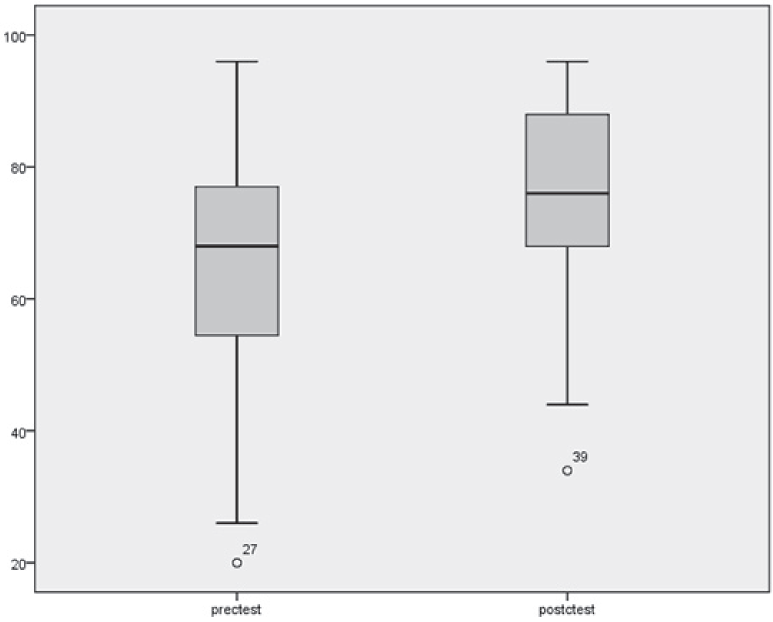

Figure 2　Boxplots of Pre-and Post-C-Tests

Before analysing the differences between these tests, the normality of these three tests is examined first. Figure 3 depicts the distribution of the mean scores of the VP 2010. Table 3 and Table 4 show the descriptive data and the results of the normality tests of the VP 2010 respectively.

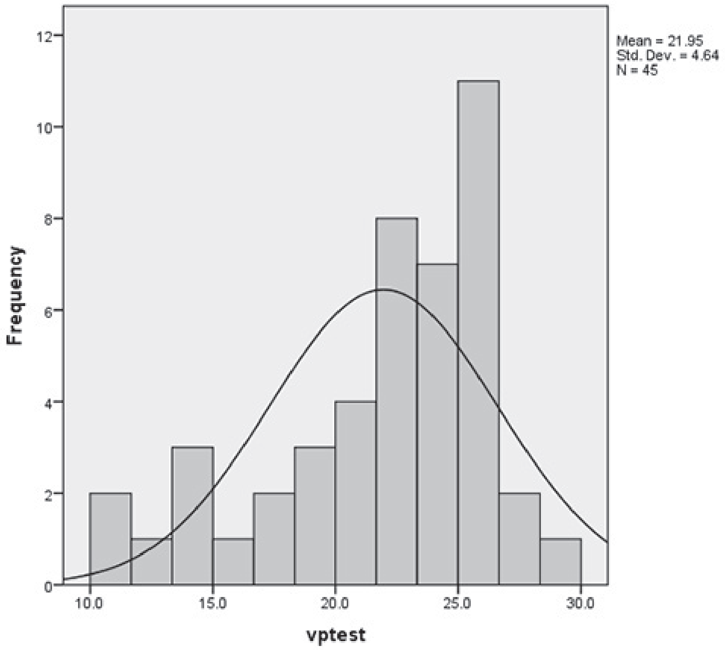

Figure 3 Histogram of the mean scores of VP 2010

Table 3 Descriptive Data of VP 2010

		Statistic	Std. Error
vptest	Skewness	−.928	.354
	Kurtosis	.011	.695

Table 4 Tests of Normality (VP 2010)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
vptest	.166	45	.003	.910	45	.002

a. Lilliefors Significance Correction

In order to identify the normality of the graph in Figure 3, the value of skewness and that of kurtosis of this data have to be checked. The value of skewness is −.928, and the standard error of skewness is .354. When the value of skewness falls within the range of normality in which the standard error is multiplied by 2 and also within plus or minus of that value, the data is normally distributed. In this case, however, the value of skewness is not included in the range of standard error of skewness (i.e., from −.708 to +.708), therefore, the distribution of skewness is not normal. Following the same process, the value of kurtosis is .011, whilst the standard error of kurtosis is .695. In addition, the value of kurtosis does not fall within the range (i.e., from −1.39 to +1.39), which indicates that the distribution of kurtosis is not normally distributed. Furthermore, the result of the Kolmogorov-Smirnov test is significant at the p=.003 level, which also shows that this data is not normally distributed and a non-parametric test can be applied to test this data.

Figure 4 is the histogram of the Pre-C-Test. This graph does not seem to be normally distributed. Table 5 represents the descriptive data and Table 6 shows the results of the normality tests of the Pre-C-Test. Examining the normality from the descriptive data, the value of skewness is −.570 and the standard error of skewness is .365, which shows that the value of skewness is within the normality range (i.e., from −.730 to +.730). In addition, the value of kurtosis is .216 and the standard error of kurtosis is .717, indicating that the distribution of kurtosis is normal (i.e., from −1.43 to +1.43). Moreover, the result of the Kolmogorov-Smirnov test is not significant (p=.200), therefore, this data is normally distributed and a parametric test can be applied to this data.

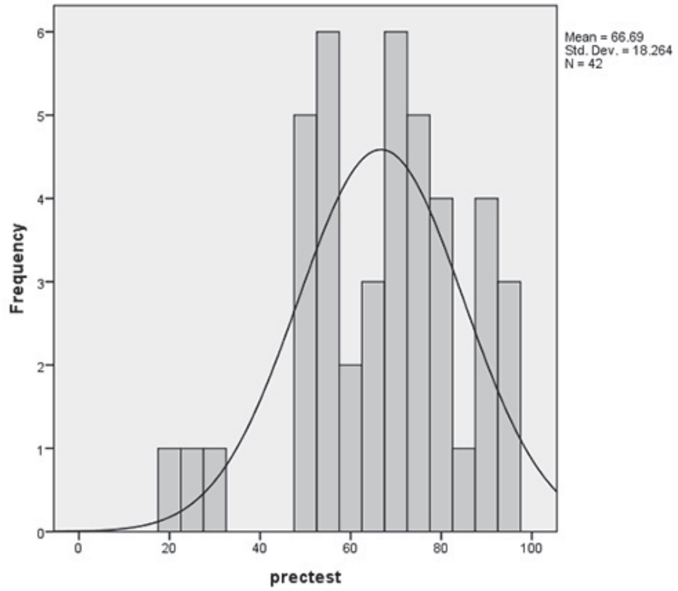

Figure 4 Histogram of the Pre-C-Test

Table 5 Descriptive Data of Pre-C-Test

		Statistic	Std. Error
prectest	Skewness	−.570	.365
	Kurtosis	.216	.717

Table 6 Tests of Normality (Pre-C-Test)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
prectest	.082	42	.200*	.961	42	.155

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Lastly, Figure 5 shows the distribution of the Post-C-Test. Table 7 and Table 8 represent the descriptive data and the results of the normality tests of the Post-C-Test. The value of skewness is −.911 whereas the standard error of skewness is .365. The distribution of skewness does not fall within the normality range (i.e., from −.730 to +.730). However, the value of kurtosis is .541 and the standard error of kurtosis is .717 (i.e., from −1.43 to +1.43), which means that the value of kurtosis is normally distributed. Furthermore, no significant differences can be found according to the Kolmogorov-Smirnov test (p=.200), which proves to be a normally distributed data.

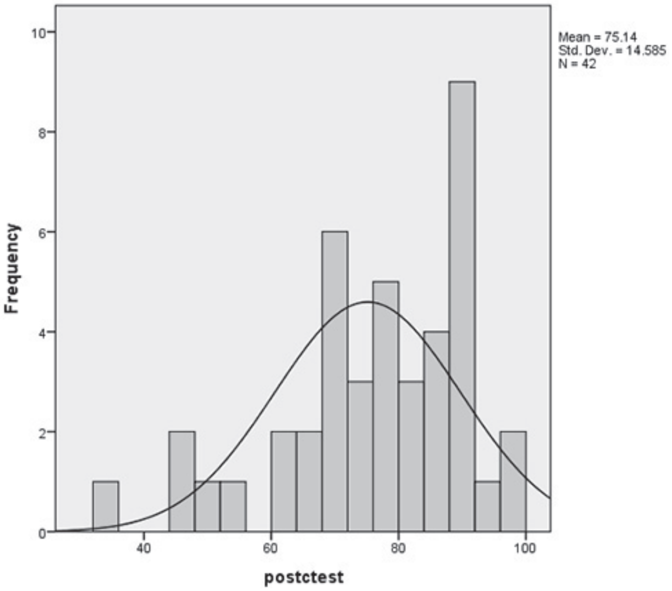

Figure 5 Histogram of the Post-C-Test

Table 7 Descriptive Data of Post-C-Test

		Statistic	Std. Error
postctest	Skewness	−.911	.365
	Kurtosis	.541	.717

Table 8 Tests of Normality (Post-C-Test)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
postctest	.109	42	.200*	.932	42	.015

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

After examining and confirming the normal distribution of each set of data, the specific type of statistical measurement can be used to analyse them. Since the VP 2010 data is not normally distributed, a non-parametric measurement is applied. On the other hand, the distributions of the two C-Tests are normal, so a parametric test is used.

In order to determine if there are statistical differences between the two C-Tests (the Pre-and Post-Tests), a paired samples t-test, which is one of the types of parametric test, is conducted. The result of this data is shown in Table 9.

Table 9 Paired Samples T-Test of Pre-and Post-C-Tests

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 prectest − postctest	−8.487	13.605	2.179	−12.897	−4.077	−3.896	38	.000

The difference between the Pre-and Post-C-Tests is found to be significant (t=-3.90, df=38, p=.001). This means that there are significant statistical differences between the two C-Tests, which means that the students made progress in their C-Test scores.

Next, in order to identify the relationship between the VP 2010 and the Pre-and Post-C-Tests, the correlation coefficients of these three scores are examined. As a measuring method, Spearman's rho, a non-parametric correlation, is used. As shown in Table 10, the correlation coefficient between the scores of the Pre-and Post-C-Tests is high (ρ = .621). Also, the correlation coefficients between the scores of the VP2010 and the two C-Tests (the Pre-and Post-Test) are .477 and .625 respectively, which show that the correlation figures became higher between the VP 2010 and the Post-C-Test. Moreover, Cronbach's alpha coefficient among the mean scores of these three tests results in an adequate level ($\alpha = .71$). These results need to be analysed in more detail to check if there are significant statistical differences among these tests.

Table 10　Correlations of Pre-and Post-C-Tests and VP 2010

	Pre-C-Test	Post-C-Test	VP 2010
Pre-C-Test	—	.621**	.477**
Post-C-Test		—	.625**
VP 2010			—

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

5. Discussion

5. 1　The Design of VP 2010

In the process of developing a test design, it is important to set up the purpose of the test. With regard to creating a vocabulary test, Read (2000) proposes three dimensions where the scope of vocabulary assessment can be explained in a contrasting way. The first dimension is concerned with the construct, or vocabulary knowledge, of the test. For measuring a distinct construct, a discrete test is used, while an embedded test is utilized for assessing broad construct such as writing ability. The second dimension is related to the assessment of the range of vocabulary. When the knowledge of the specific word items is assessed, it is a selective vocabulary measure. On the other hand, when the whole word items of the materials are measured, it is a comprehensive one. Lastly, the last dimension is connected to the role of context. If the vocabulary knowledge is assessed with no reference to the context in the test, it is context-independent. If the test taker's ability to produce words is assessed with reference to the context, it is context-dependent. When the concept of three dimensions is applied to the VP 2010, the revised version is classified as discrete and selective, in that the vocabulary knowledge is measured separately from other components of language ability, and its assessment is based on a set of selected target words from the defining vocabulary. As for the last dimension, it should be relatively context-dependent. Although there is a debate as to whether cloze tests are classified as context-independent or context-dependent (Read 2000: 12), the degree of context dependence is determined by the design of the test itself.

The VP 2010 is designed in such a way that the test takers are expected to complete the missing letters referring to the context as a clue, which would make the test more context dependent.

As for the evaluation of the vocabulary tests, Nation (2008) presents the reliability as one of the features of good tests. According to Nation (2008: 153), the details of the conditions of reliability are as follows: (a) the test contains at least 30 items or points of assessment; (b) the test format is familiar to the learners because they have taken such a test before; (c) the instructions and way of answering are the same in all versions of the test; (d) the marking uses a marking key and criteria that take account of the most possible variations in answering. When applied to the VP 2010, each test sheet has 30 word items to be assessed. Since the sample test is presented to the students before taking the test, the test format is familiar to them. Moreover, the same type of test is conducted to test the 30 word items every week. Lastly, the VP 2010 is designed with only one possible correct answer by referring to the initial letters of the word and the context, so it does not allow any alternative response. In terms of fulfilling the conditions above, it can be said that the VP 2010 is reliable when it is properly used for its purpose.

5. 2 Findings from Statistic Results

The three tests (VP 2010, Pre-and Post-C-Tests) are examined and statistic results are obtained in section 4. As for the VP 2010, considering the fact that Cronbach's alpha coefficient among the mean scores of the VP 2010 is quite high (α = .97), it seems possible that each of the 10 test scores is consistent. Comparing the Pre-and Post-C-Tests, the mean score of the Post-C-Test increases (i.e., from 66.7 to 75.1), and also the results of the standard deviation show that the distribution of the scores is more concentrated around the mean score in the Post-C-Test (i.e., from 18.26 to 14.59).

With regard to research question (1), the results of a paired samples t-test indicate that there is a significant difference between the Pre-and Post-C-Tests. This result may be explained by the fact that after working on the vocabulary tests for 10 weeks, the participants became accustomed to the tests in which they have to use the context as a clue. Moreover, the participants might have learned to pay more attention to the inflectional or derivational forms of words with the VP 2010. Therefore, the participants in this study made an improvement in their scores of the Post-C-Tests.

Another important finding, regarding research question (2), is that the correlations among these three tests (VP 2010, Pre-and Post-C-Tests) indicate moderately high correlations. A possible explanation for this might be that the C-Tests and the VP 2010 are similar in their test format compared to the former version, VP 2008. Both test types are designed to measure the controlled productive ability of vocabulary. However, correlations obtained from these tests are not very high, though not too low. The reason for this is that whilst the participants are encouraged to study the test items and prepare for the vocabulary test every week, there is no time allowed for preparing for the C-Tests; also the participants need to respond using their whole word knowledge, which makes the C-tests more difficult and challenging tests.

Judging from these findings, the revised VP 2010 is reliable in that each test has consistent results and there are positive correlations between the Pre-and Post-C-Tests, and lastly, the scores of the Post-C-Tests improved significantly.

6. Conclusion

The present study is designed to examine the reliability of the revised vocabulary practice tests. After reviewing the previous study on the VP 2008, the process of development of the revised VP 2010 is presented. Then, from the viewpoint of design, the VP 2010 is a more context-dependent test to assess learners' productive word knowledge. Moreover, by examining and analysing the statistic data from the VP 2010 and the C-Tests, the study has shown that there are significant differences among these three tests, and the effects of the VP 2010 are reflected in the learners' score improvement in the Post-C-Test. Taken together, these results suggest that the VP 2010 is a reliable vocabulary instruction tool, which will encourage the learners to study the defining vocabulary and also develop their productive word knowledge.

The findings in this study are subject to the following two limitations. First, as Nation (2008) suggests, a good vocabulary test has to meet the requirement of validity, and practicality along with reliability. Considering the fact that it does not take long to take the test, it is easy to mark, and the score is easy to interpret, the VP 2010 is a practical test. In addition, in order to fulfill the condition of validity, careful analyses are required. Since reliability and validity are dependent on the context in which a test is used, closer examination of the VP 2010 is necessary. Another limitation is the number of participants. In this study, the number of participants is 45, which is rather small for analysing the data. If there were more participants, the results of the correlations could be affected, and there would be more opportunities to prove the strong relationship between the test scores.

In order to examine the VP 2010 in more detail, further research needs to be done to analyse the productive knowledge of the participants. For example, by collecting their writing samples and analysing the lexical diversity in their writing, a significant difference might be found between the tests and the writing samples. In addition, it would be useful to compare the results of the vocabulary practice tests with those of such established tests as the Vocabulary Levels Test. Since such kind of tests are widely used in the ESL context and fulfill the conditions of reliability and validity, more precise results would be available in analysing the data. Moreover, it is important to conduct a questionnaire on the vocabulary practice activities with the series of word lists. Examining the learners' responses, their attitudes towards vocabulary learning would be revealed. For instance, it would be useful to check whether they used the monolingual English dictionary when they prepared for the vocabulary practice test every week. The findings might be one of the factors that could affect the improvement of the learners' vocabulary knowledge. Also, those findings would be of great help to devise the method of vocabulary instruction in class. Through monitoring the learners' improvement, the instructors give the learners incentives to keep studying on a regular basis. This may well contribute to the further development of the vocabulary instruction materials in the future.

References

Fraenkel, J. R., and Wallen, N. E. (1993). *How to design and evaluate research in education*. 2nd edition. New York: McGraw-Hill Inc.

Kinshi, K. (2009). Bringing vocabulary practice into active use: Towards motivated learning. *The Bulletin of the Writing Research Group, JACET Kansai Chapter*, 8, 15-26.

Kinshi, K. (2011). A Questionnaire Survey Based on Vocabulary Practice Using English Monolingual Dictionaries. *The Bulletin of the Writing Research Group, JACET Kansai Chapter*, 9, 25-37.

Matsui, S., Okada, T., Ishihara, K., & Pavloska, S. (2004). Toward the use of monolingual English dictionaries: Building knowledge of defining vocabulary. *Doshisha Studies in Language and Culture*, 7, 83-112.

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

Nation, I. S. P. (2008). *Teaching Vocabulary: Strategies and Techniques*. Boston: Heinle.

Okada, T., Ishihara, K., Kinshi, K., & Pavloska, S. (2009). Practice Activities with Defining Vocabulary: Making English Dictionary Entries More Accessible. *Doshisha Studies in Language and Culture*, 12, 1-38.

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

Writing Research Group, JACET Kansai Chapter. (Ed.). (1995). Teaching writing in colleges and universities: A survey report.

Writing Research Group, JACET Kansai Chapter. (Ed.). (1998). Teaching writing in colleges and universities: Practical report.