Vocabulary Tests for Junior High School Students

David Aline
Kanagawa University

This study examines the vocabulary levels of Japanese junior high school students through the use of vocabulary recognition tests. Vocabulary recognition tests measure language in use rather than only decontextualized vocabulary items. The primary objective of this study was to analyze the equivalency across modalities and across forms of a vocabulary recognition test. The secondary purpose of the study was to look at the vocabulary level of Japanese junior high school students on recognition tests.

本研究では、語彙認知テストを通じて日本の中学生の英語語彙認知能力を考察した。語彙能力テストで出題される語彙は文脈から外れた語彙でなく、文脈中の語彙である。この研究では、二つの形式とモダリティーの違う語彙認知テストの均等性を調べることを主の目的とした。また、語彙認知テストにより日本の中学生の英語語彙レベルを検証することを当研究のもう一つの目的とした。

Testing students' vocabulary is a useful way of ensuring that students are on track in acquiring the vocabulary being taught. A test can also be a motivating experience, as students will have tangible evidence of achievement in the form of test scores. Nation (1990, p. 8) lists six reasons for testing vocabulary:

1. To find learners' total vocabulary size.
2. To compare vocabulary knowledge before and after the course.
3. To keep a continuing check on progress.
4. To encourage learning by setting short-term goals.
5. To see the effectiveness of your teaching.
6. To investigate learning.

After determining the reason or reasons for giving a vocabulary test, the teacher or administrator will need to settle on the type of test to be employed. Two questions that arise when deciding on the type of vocabulary test to use in assessing students' current levels or their gains over time are: 1) what vocabulary will be tested, and 2) how will the vocabulary be tested.

In deciding the vocabulary to be tested, the teacher can consider for examination the vocabulary directly taught in class, the vocabulary contained in the textbook (often listed in the appendices), the vocabulary required to be learned by the students by the Japanese Ministry of Education (Mombusho), or the most frequent vocabulary in English as listed by frequency counts (e.g., Thorndike and Lorge, 1944; West, 1953). The choice will depend on how the results are to be applied. Testing vocabulary taught in class could be used to analyze teaching methodology or technique. Testing vocabulary learned from the textbook might help in textbook selection or in review of units in the book that have not been learned by the students. Testing required vocabulary will ensure that the students are prepared for official examinations. In addition, testing high-frequency vocabulary might help in selecting reading or listening materials at the appropriate level for maximum comprehension by the students.

The second question asks how the vocabulary will be tested. Vocabulary can be tested as recognition vocabulary or recall vocabulary (Nation, 1990, p. 79). A vocabulary test of recognition assesses the students' ability to show recognition of the meaning of an English word through translation, definition, or matching the meaning to a picture. A recall test supplies the translation, definition, or picture, and the student must furnish the correct English word. A variety of testing techniques is utilized for both types of tests (see Nation, 1983, 1990, for a complete discussion of vocabulary testing).

Testing can also depend on modality. Recall can be tested through the written or spoken modes, a recognition test can make assessment by way of reading or listening modalities. Again, the method of testing will depend on how the results will be used.

This study looks at one type of vocabulary test across modalities.

The Study

The primary objective of this study was to analyze the equivalency of two forms of a vocabulary recognition test and to analyze the forms across modalities. The secondary purpose of the study was to look at the vocabulary recognition levels of Japanese junior high school students.

The research questions examined were:

1. Are the two forms (A and B) of the 1,000-word level vocabulary test equivalent?
2. Are the two forms (A and B) of the 1,000-word level vocabulary test equivalent across the modalities of reading and listening?
3. What are the vocabulary levels of intact classes of Japanese junior high school students?

In order to examine these questions statistically, two null hypotheses were formulated to test the two research questions. The null hypotheses for questions 1 and 2 were, respectively:

1. There is no significant difference between the means for Form A and Form B across the same modality of reading.
2. There is no significant difference between the means for Form A and Form B across the modalities of reading and listening. Additionally,

to determine that no difference existed across the two modality type groups, a third hypothesis was tested:

3. There is no significant difference between the group taking the test across the same modality, reading, and the group taking the test across different modalities, reading and listening.

Participants

The subjects participating in this study were 94 Japanese junior high school students at a private school in Tokyo. The students were in two intact classes in their second year of junior high school. All subjects were males with an average age of 13 to 14. Each of the two classes had 47 students. The figures of 94 subjects with 47 in each class are for subjects who completed both forms of the test; three subjects were excluded as they took only one form because of absence on the day of test administration. Unfortunately, because of curriculum constraints within the institution, the subjects could not be randomized into the two groups. Therefore, since it is important to demonstrate that the two groups are performing the same across similar forms of the test, hypothesis number 3 was tested.

Materials

The materials in this study consisted of Forms A and B of the 1,000-Word Level Vocabulary Test designed

by Nation (1995). The test measures the first 1000 words of the General Service List (West, 1953). The two forms, each consisting of 40 items, assess the ability of the student to recognize the vocabulary item within the context of a sentence. This type of test measures language in use rather than only recognition of decontextualized vocabulary items. Therefore, it is a better measure of the student's ability to understand a word within a meaningful context. The student reads the sentence containing the item to be tested and marks in the blank with a "T if a sentence is true...N if it is not true...X if you do not understand the sentence" (Nation, 1995, p. 40-41, test instructions). For the population of students in this study, the instructions were translated into Japanese and the marks changed for the Japanese context. Thus, the subjects were to mark O if the sentence was true, X if the sentence was false, and ? if they did not understand. Each vocabulary item is tested twice across the two forms but within different contexts. For example, the first item on Form A is, "We cut time into minutes, hours and days" (p. 40). The vocabulary item tested is "time". This item is then tested in a different context on Form B in the test question, "We can stop time" (p. 41). Some of the items include pictures with the sentences to clarify the statement. Form A uses 7 pictures; Form B uses 6 pictures.

For scoring, a single form can be scored as a true-false test, but this means that the students have a 50% chance of guessing the item correctly, which causes a 40-item true-false test to become a 20-item test when guessing is factored out. Therefore, it is suggested that the two forms could be used together and the item only counted as correct if the items on both tests are answered correctly. This procedure scores the test as a multiple-choice test since the range of possible guesses is a ratio of 1:4. The reliability of the test is consequently increased. The administration and correcting time of the test is also increased. The choice of scoring technique will depend on the time available and the use to which the scores will be put.

The reading modality used forms A and B in their reading forms. To test across modalities, Form A was used in its reading format and Form B was altered. For Form B in the listening modality, the written statements were removed from the scoring sheet and only the blanks, pictures, and Japanese instructions were left intact. Item numbers were of course included to help the subjects understand which item they were on.

Procedures

The two intact classes sat the tests at the end of the spring semester in July. The forms were administered during the regular class period. Group 1 completed Form B in its regular reading form and Group 2 completed Form A also in its regular reading form. The students were informed that the scores would not affect

their grades but would be used to assist the teacher in preparing materials at their vocabulary level. After the summer vacation, two weeks into the autumn semester, the students took the second form of the test. Group 1 took Form A in its regular reading format and Group 2 took Form B in its revised listening format. This counterbalanced research design is shown in Table 1.

Table 1: *Counterbalanced Research Design*

Group	Test Forms	
Group 1	Reading Form B	Reading Form A
Group 2	Reading Form A	Listening Form B

In the administration of the listening format for Form B to Group 2, each item was read twice by the regular classroom teacher. The teacher read the items at what she judged to be an appropriate listening speed for her students.

The students took a little more than 15 minutes to complete the forms in the reading modality and just under 20 minutes to complete Form B in the listening modality. The teacher reported that the students appeared to have more confidence in the reading modality than in the listening modality since they have had extensive experience with written tests but not with listening tests.

Analyses

The independent variables in this study are the modality groups, reading only versus reading and listening, and the test forms, A and B. The dependent variables are the test scores. Therefore, statistical analysis consisted of a Multiple Analysis of Variance (MANOVA) used to determine the overall mean comparisons. The MANOVA was used to ensure that the groups were the same since they could not be randomized. To determine the equivalency of the test forms, the means were compared and separate matched-pairs t-tests for repeated-measures design were performed across the forms. Pearson product-moment correlation determined the strength of the relationship, and the reliability of the forms was tested using Cronbach alpha. All of the assumptions for the statistics used in this study were checked and met.

Results

The results of the MANOVA showed no significant difference between the two class groups as can be seen in Table 2. This shows that the two classes were not different. Furthermore, there was no interaction effect to be explained.

Table 2: *Univariate F-tests with (1,92) D.F. on Group by Test Form*

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
Form A	11.585	1526.85	11.585	16.596	.698	.406
Form B	.862	1925.57	.862	20.93	.041	.840

It is unclear from the MANOVA what the differences are between the modalities on the forms and between the forms themselves. Therefore, the means were compared and independent matched-pairs t-tests were performed on Forms A and B for the reading modality, and between Forms A and B for the reading by listening modalities. Table 3 reports the descriptive statistics for the test forms and modalities.

Table 3: *Descriptive Statistics for Test Forms across Modalities*

	Mean	Std Dev	Min	Max	N
Form A reading	24.26	3.97	18	37	47
Form B reading	23.49	5.08	11	34	47
Form A reading	25.23	3.85	17	32	47
Form B listening	23.30	4.01	13	32	47

Since a comparison of the means between the forms within each modality appeared to show similarity, independent t-tests were completed for each (Table 4). The level of significance was set at $p<.05$. However, since two t-tests were to be performed on the same data set, the significance level was halved, divided by 2, to take into consideration the increased probability of finding a spurious result. The revised significance level was set at $p<.025$

Table 4: *Independent Matched-Pairs T-Tests Within and Across Modalities*

Forms	P
Forms A and B reading	.258
Form A reading and Form B listening	.014*

$p<.025$

No significant difference was found between Forms A and B when administered within the same modality, reading. Therefore, it is possible to say that these forms are not different and can be used for looking at students' improvement in vocabulary over time. Across the modalities of reading and listening for Forms A and B, respectively, there was a significant difference between the two forms at $p<.014$. The forms are therefore different and could not be used in comparison of improvement over time when modalities are crossed

by the forms. That is to say, students' scores on Form A in the reading modality could not be compared with students' scores on Form B in the listening modality. However, Form B in the reading modality and Form B in the listening modality were not significantly different as shown in the MANOVA in Table 1. Therefore, the same form could be used across modalities if enough time elapsed that the students would have forgotten the questions on the test. However, this is testing between groups and the forms should be checked for equivalency within groups.

The strength of the relationship for the two forms within the same modality of reading was calculated by the researcher with Pearson product-moment correlation. The correlation coefficient was significant at .5092, which shows a related variance of 26%. This is not a great amount of variance if the test forms are to be used for research purposes, but might suffice for a teacher's classroom use in looking at vocabulary acquisition.

The reliability coefficients for the forms within each modality as used in this study were calculated with Cronbach alpha and are reported in Table 5.

Table 5: *Reliability Coefficients for Each Form Administration*

Forms by modality	Á
Form A reading	.5046
Form B reading	.6997
Form A reading	.5794
Form B listening	.4831

The reliability coefficients for these administrations are not very high, especially for the listening modality. When both groups and the forms were combined, the reliability was .7140. These reliability coefficients are reported for the test forms as scored on a true-false basis. If both Forms A and B are scored as one test, they can be scored as a multiple-choice test. This improves the reliability of the instrument, but precludes use of different forms for looking at vocabulary gains over time.

The average scores for the students are reported in Table 3. Most of the students received a total score between 23 and 25 points out of a possible 40 test items. Their vocabulary level according to these tests is therefore a little over half of the first 1,000 most frequent words of English for recognition in context. However, when scored as a multiple choice test, a correct point is awarded only when both items testing the same

word on the different forms are correct, the average score dropped to 15.25 (15.02 for reading modality, 15.47 for reading and listening modalities). The drop is not dramatic but gives a more accurate picture of the students' vocabulary level.

Discussion

The results show the two forms to be not different when used in the reading modality. Therefore, they could be utilized by teachers within the classroom for most of the six reasons for testing vocabulary outlined in the introduction of this study. However, caution must be taken when the test is used to investigate learning. Use of the test will of course depend on the research questions to be answered.

The students' scores show that they are still within the area of the highest frequency words of English. High frequency words should be directly taught by the teacher because they give a high return for the time and energy invested in learning. These words will be met very often, will be met soon so that they will be reinforced in learning, and will help to increase the basis of comprehension by quickly raising the level of percentage of words known in a text. Knowledge of high frequency words will also raise the students' confidence as they see that learning quickly pays off, thus lowering the affective filter.

Further research in this area could use the forms to compare vocabulary acquisition rates with classroom interaction within different strand: meaning focused input, form focused input, meaning focused output, and fluency practice. The forms also need to be checked for equivalency within the same modality of listening. Results of the test could also be compared with other tests, i.e.: proficiency level, reading ability, oral interview, etc. These research areas will ensure the reliability and validity of the test.

Conclusion

The two forms of the 1,000-word level vocabulary test were found to have no difference within the same modality, but were different when tested across modalities. When used in the reading modality, the forms of this test can safely be used for analyzing students' vocabulary improvement over time.

References

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

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5(1), 12-25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston: Heinle & Heinle Publishers.

Nation, I. S. P. (1995). *Teaching and learning vocabulary*. Unpublished manuscript; Course workbook. Japan: Temple University Japan.

Thorndike, E., & Lorge, I. (1944). *The teacher's word book of 30,000 word*. New York: Teachers College, Columbia University.

West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.