MENDIDIK: Jurnal Kajian Pendidikan dan Pengajaran
Volume 1, No. 2, Oktober 2015: Page 163-170
ISSN: 2443-1435

EVALUATING THE VOCABULARY TEST

Nia Pujiawati[1]

ABSTRACT: This study describes the analysis of vocabulary test given to the students seen from the difficulty level (ID) and discriminating power (DP) point of view and also from reliability and validity judgment. This vocabulary test was conducted at SDN Puseurjaya, an elementary school located at Karawang, from January 30th 2012 through February 1st 2012. After doing the analysis on the items, it was found that the test of vocabulary given to the second grade of SDN Puseurjaya students meets almost all requirements to be the acceptable test items. However, as it does not reach validity, the writer then needs to check and review in depth to make some improvements for the next test construction.

Keywords: Evaluating, Vocabulary Test.

INTRODUCTION

English has been considered to be the first foreign language in Indonesia. It functions to help the development of the state and nation, to build relations with other nations, and to run foreign policy including as a language used for wider communication in international forum. In relation to that Indonesia has been carrying out teaching EFL in almost level of schools, including elementary. At the elementary school, English is taught as local content subject in which the students are expected to have skills of the language in simple English with emphasis on listening, speaking, reading, and writing skills using selected topic that relates to their environmental needs.

In the context of learning English as a foreign language, vocabulary is one of fundamental and important component that elementary learners should become skilled at. Vocabulary simply refers to the knowledge of words and word meaning that relates to the four basic English skills whether it is receptive (reading and listening) or productive (speaking and writing). As Rivers in Nunan (1991: 117) argues that the acquisition of an adequate vocabulary is essential for successful second language use because without an extensive vocabulary, we will be unable to use the structures and functions we may have learned for comprehensible communication. It means the students who don't have large vocabularies will often struggle to achieve comprehension even it is a very simple one, and to make it worse – at least based on the writer experience as a teacher – they will become so frustrated that make them find English more difficult to learn than they thought before. In the other way around, when students have a strong vocabulary, they will be more confident in receiving or producing the skills.

Since it plays a very important role in English learning as described above, it automatically takes important part in assessment. The teacher should regularly give the test to the students to get the information how well the

[1] Program Studi Pendidikan Bahasa Inggris FKIP Universitas Singaperbangsa Karawang; nioth_euy@yahoo.com

students acquire the materials at the desired level and also to measure whether the students have achieved the learning objectives that have been formerly targeted. The most important thing, the test done by the teacher will not only tell the students' progress but also allow the teachers to make adjustments and improvements to the instruction applied in the classroom to meet the needs of their students.

However, as the test given to the students is a teacher-made test that is not commercially produced and standardized, which means that the quality of the test is questionable and open to debate, it needs an investigation to see whether or not the writer - in this case the teacher – design the effective and meaningful test. On this basis, this paper then describes the analysis of vocabulary test given to the students seen from the difficulty level (ID) and discriminating power (DP) point of view and also from reliability and validity judgment.

METHODS

This vocabulary test was conducted at SDN Puseurjaya, an elementary school located at Karawang, from January 30th 2012 through February 1st 2012. Since the time is very limited, the test was primarily given only to the class A of second grade consisting twenty-nine students. However, to get precise data that may support the possible result, the writer then decided to invite eleven students of class B of the same grade to take part in the test. Those students were chosen randomly. In other words, the writer did not take any factors into consideration in selecting those additional test respondents.

Initially, the writer constructed the test to be delivered to the students. The test which consists of fifty items cover a variety of test format including thirty multiple choice items, ten matching items, and ten fill-in-the-blank items. The reason behind choosing this various type of questions was just to avoid the students from being bored stiff and tired of doing the unchanged style test from number one up to number fifty. As it is mentioned above, the test was designed to see the students progress on mastery of vocabulary, taught as part of English as a local content subject; therefore, the writer believes that the materials being tested are correlated to course objectives and learning standards.

Having constructed the test, the writer then distributed the test to the students to try it out. The results of students' performance in this test were then used to determine the difficulty level (ID) and discriminating power (DP) of each item that will be discussed respectively hereafter. After getting the ID and DP, the writer continued to retest the test to the same group of students with different amount of questions and on different day. The result then again was analyzed to see whether the test meet reliability and validity.

DISCUSSION

The difficulty level (ID) is understood as the proportion of the persons who answer a test item correctly. To calculate the difficulty of an item, the number of persons who answered it correctly is divided by the total number of the persons who answered it as we can see in the following formula:

$$ID = \frac{RU + RL}{N}$$

To perform this item analysis, the respondents' tests were arranged in order from the one with the highest cumulative score to the one with the lowest score. Then, 27.5% from the higher and 27.5% from the lower groups were taken for the purpose of comparison. In this study context, by following the formulation above the total number of respondents included in the item analysis is 11 from the upper group and 11 from the lower group; the total was twenty-two. After that, each item was analyzed using Baker criteria that suggests the item which have ID ranging from .25 to .75 can be included in the test which means the difficulty level is good. The complete computation of ID can be seen below, combined with the discriminating power.

Discriminating power (DP) is another item analysis that has the same importance as difficulty level. It is considered as the basic indicator of an item's quality; it tells those who do well on the test and those who do poorly. The discriminating power can be measured by comparing the number of students with high test scores who answered that item correctly with the number of students with low scores who answered the same item correctly, with the formula as follows:

$$DP = \frac{RU - RL}{5N}$$

Applying this formula, the writer to begin with did the same thing as it was in calculating ID. It means the respondents' tests were arranged in order from the one with the highest cumulative score to the one with the lowest score. Then, 27.5% from the higher and 27.5% from the lower groups were taken for the purpose of comparison. After that, each item was analyzed using Henning criteria that suggests the item which have DP between 0.33 to 0.67 can be included in the test which means the discriminating power is good. To make it clearer, the following Table 1 describes how the formula of ID and DP work.

From the Table 1, it is obvious that there are thirty-one items out of fifty which meet the difficulty level criteria suggested by Baker, ranging from .27 to .72. In other words the difficulty level of the vocabulary test given to the second grade students reached 62 percent. In terms of discriminating power, from the table above it is found that there are thirty items that meet Henning criteria ranging from .36 to .63, and twenty items of them did not. It means 60 percent of the total items are considered good. Then, if it is seen from both sides, difficulty level and discriminating power, there are only twenty-three items that up to standard. It can be said that 46% of the items are consider to have good quality; therefore, they are acceptable to be included in the test.

TABLE 1. Difficulty Level (ID) and Discriminating Power (DP)

No of Item	RU	RL	RU+RL	RU-RL	ID	DP
1	11	9	20	2	0.90909091	0.18181818
2	10	9	19	1	0.86363636	0.09090909
3	10	5	15	5	0.68181818	0.45454545
4	11	6	17	5	0.77272727	0.45454545
5	11	6	17	5	0.77272727	0.45454545
6	6	3	9	3	0.40909091	0.27272727
7	3	1	4	2	0.18181818	0.18181818
8	11	7	18	4	0.81818182	0.36363636
9	11	8	19	3	0.86363636	0.27272727
10	1	2	3	-1	0.13636364	-0.0909091
11	10	3	13	7	0.59090909	0.63636364
12	11	5	16	6	0.72727273	0.54545455
13	6	2	8	4	0.36363636	0.36363636
14	10	6	16	4	0.72727273	0.36363636
15	11	9	20	2	0.90909091	0.18181818
16	11	4	15	7	0.68181818	0.63636364
17	8	3	11	5	0.5	0.45454545
18	11	6	17	5	0.77272727	0.45454545
19	10	5	15	5	0.68181818	0.45454545
20	9	5	14	4	0.63636364	0.36363636
21	11	5	16	6	0.72727273	0.54545455
22	11	7	18	4	0.81818182	0.36363636
23	11	5	16	6	0.72727273	0.54545455
24	4	1	5	3	0.22727273	0.27272727
25	7	3	10	4	0.45454545	0.36363636
26	11	3	14	8	0.63636364	0.72727273
27	11	4	15	7	0.68181818	0.63636364
28	7	2	9	5	0.40909091	0.45454545
29	6	1	7	5	0.31818182	0.45454545
30	6	2	8	4	0.36363636	0.36363636
31	4	1	5	3	0.22727273	0.27272727
32	11	2	13	9	0.59090909	0.81818182
33	11	6	17	5	0.77272727	0.45454545
34	11	8	19	3	0.86363636	0.27272727
35	6	3	9	3	0.40909091	0.27272727
36	11	7	18	4	0.81818182	0.36363636
37	2	0	2	2	0.09090909	0.18181818
38	5	1	6	4	0.27272727	0.36363636
39	10	2	12	8	0.54545455	0.72727273
40	2	1	3	1	0.13636364	0.09090909
41	11	9	20	2	0.90909091	0.18181818
42	0	0	0	0	0	0
43	11	5	16	6	0.72727273	0.54545455
44	9	2	11	7	0.5	0.63636364
45	11	5	16	6	0.72727273	0.54545455
46	8	4	12	4	0.54545455	0.36363636
47	10	0	10	10	0.45454545	0.90909091
48	10	0	10	10	0.45454545	0.90909091
49	5	1	6	4	0.27272727	0.36363636
50	6	2	8	4	0.36363636	0.36363636

Since the items that may be included in the test are almost half of the full amount, the writer continued the analysis of test instrument by retesting the acceptable test to the students to see its reliability and validity, as showed at Table 2.

TABLE 2. Students Scores and Reliability Coefficient

No	Students Code	X	Y	Rx	Ry	D	D^
1	EVT01	74	82.6087	6.5	9	-2.5	6.3
2	EVT02	88	91.30435	1	2.5	-1.5	2.3
3	EVT03	34	56.52174	38.5	30	8.5	72.3
4	EVT04	70	86.95652	11	5.5	5.5	30.3
5	EVT05	42	34.78261	34.5	37.5	-3	9.0
6	EVT06	70	69.56522	11	17.5	-6.5	42.3
7	EVT07	66	60.86957	15	28	-13	169.0
8	EVT08	72	65.21739	8.5	23.5	-15	225.0
9	EVT09	70	78.26087	11	12.5	-1.5	2.3
10	EVT10	64	86.95652	16.5	5.5	11	121.0
11	EVT11	56	65.21739	23.5	23.5	0	0.0
12	EVT12	78	86.95652	4	5.5	-1.5	2.3
13	EVT13	52	65.21739	26	23.5	2.5	6.3
14	EVT14	64	39.13043	16.5	36	-19.5	380.3
15	EVT15	48	86.95652	30	5.5	24.5	600.3
16	EVT16	56	52.17391	23.5	32.5	-9	81.0
17	EVT17	26	17.3913	40	40	0	0.0
18	EVT18	50	69.56522	27.5	17.5	10	100.0
19	EVT19	84	95.65217	2.5	1	1.5	2.3
20	EVT20	68	65.21739	13.5	23.5	-10	100.0
21	EVT21	40	52.17391	36	32.5	3.5	12.3
22	EVT22	72	78.26087	8.5	12.5	-4	16.0
23	EVT23	58	65.21739	21	23.5	-2.5	6.3
24	EVT24	60	56.52174	19.5	30	-10.5	110.3
25	EVT25	42	43.47826	34.5	34.5	0	0.0
26	EVT26	68	82.6087	13.5	9	4.5	20.3
27	EVT27	76	69.56522	5	17.5	-12.5	156.3
28	EVT28	44	34.78261	33	37.5	-4.5	20.3
29	EVT29	74	82.6087	6.5	9	-2.5	6.3
30	EVT30	34	69.56522	38.5	17.5	21	441.0
31	EVT31	56	56.52174	23.5	30	-6.5	42.3
32	EVT32	46	73.91304	32	15	17	289.0
33	EVT33	60	65.21739	19.5	23.5	-4	16.0
34	EVT34	48	78.26087	30	12.5	17.5	306.3
35	EVT35	50	43.47826	27.5	34.5	-7	49.0
36	EVT36	84	91.30435	2.5	2.53	-0.03	0.0
37	EVT37	56	65.21739	23.5	23.5	0	0.0
38	EVT38	64	78.26087	16	12.5	3.5	12.3
39	EVT39	48	65.21739	30	23.5	6.5	42.3
40	EVT40	36	21.73913	37	39	-2	4.0
Total						ΣD^	3501.5

In regard to reliability, the writer applied test-retest method. The writer compiled the good items that meet the requirements of difficulty level and discriminating power and then distributed to the same students. Having checked the students' performance on the test, the writer then analyzed using the Spearman rho:

$$r_s = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

The following table shows students scores and reliability coefficient in two tests using the above formula.

$$r_s = 1 – \frac{6(3501.5)}{40^3 – 40}$$

$$r_s = 1 – \frac{21009.0}{64000 - 40}$$

$$r_s = 1 – \frac{21009.0}{63960}$$

$$r_s = 1 – 0.328470$$

$$r_s = 0.671529.$$

Following the Spearman Rank Correlation Coefficient to calculate correlation (to know if two variables are related to each other), the value obtained is 0.671529 (rounded up this becomes 0.67). According to Choudhury (2009), in general $r_s > 0$ implies positive agreement among ranks, $r_s < 0$ implies negative agreement (or agreement in the reverse direction), $r_s = 0$ implies no agreement. Based on the description above, it can be interpreted that the students who got high score in the first test also got high score in the second test, and those who got low score in the first test also got low score in the second test. This shows that their responses are reliable. Thus, the measuring instrument is reliable.

Dealing with validity, the writer used content validity in this analysis. It means, as cited in Hartoyo (2011: 137), that the test assesses the course content and outcomes using formats familiar to the students. Another way of saying this is that content validity concerns, primarily, the adequacy with which the test items adequately and representatively sample the content area to be measured. Thus, when a test has content validity, the items on the test represent the entire range or larger domain of possible items that the test should cover.

To know whether the test items represent the domain or universe of the trait or property being measured, the writer identified the overall content of the test to be represented by using curriculum of English at elementary school. It is stated that the students of second grade will learn about self-introduction, family, parts of body, clothes, numbers, days, meals, and doing things/activities. Thus, the questions presented in the test should be about vocabularies related to those topics.

Having analyzed the test items given to the students, the writer found there are fourteen out of fifty items that correspond to content pointed out in the curriculum as we can see Table 3.

TABLE 3. Test Items

No.	Content subject	Total
1	Self-introduction	-
2	Family	1
3	Parts of body	1
4	Clothes	2
5	Numbers	5
6	Days	2
7	Meals	1
8	Doing things/Activities	2

The table above shows that it only represents 30% of the content area to be measured. The other 70% of the test surprisingly represents the materials that should be learnt in first grade and even third grade. Because of the percentage between the test items and the content area is very low; thus, the test instrument can be assumed not valid.

CONCLUSION

English vocabulary is one of the elements in teaching English at the elementary school. It plays important role in determining the successful of the students in learning the English language skills such as reading, listening, writing, and speaking though it is in a very simple context. As a result, it plays important part also in assessment. Since the quality of the test that made by the teacher is questionable, it should be investigated to see the level of its difficulty, discriminating power, reliability, and validity.

After doing the analysis on those items, it was found that in terms of difficulty level the test items reached 62%, and 60% dealing with the discriminating power. These results denote that the test have good quality. Then, if the test is analyzed using both difficulty level and discriminating power, there are only twenty-three items that can be included in the test. Moreover, seen from the reliability and validity standpoint, which was analyzed using Spearman rho and content validity, the test is one hand considered reliable, but on the other hand it is considered not valid.

In conclusion, the test of vocabulary given to the second grade of SDN Puseurjaya students meets almost all requirements to be the acceptable test items. However, as it does not reach validity, the writer then needs to check and review in depth to make some improvements for the next test construction.

REFERENCES

Choudhury, A. (2009). *Spearman Rank Correlation Coefficient*. Retrieved on February 3rd, 2012, from http://www.experiment-resources.com/spearman-rank-correlation-coefficient.html.

Hartoyo. (2011). *Language Assessment*. Semarang: Pelita Insani.

Nuna, D. (1991). *Language Teaching Methodology: A textbook for teachers.* Sydney: Prentice Hall International (UK) Ltd.