CHAPTER 5

CONCLUSIONS AND SUGGESTIONS

This chapter deals with the conclusions and the suggestions based on the results and the discussions of this research.

5.1. Conclusions

The findings of the research specify that not all items in the final semester test have good validity, in relation to construct validity, content validity, and face validity. The construct validity and the content validity of the final semester test are valid, except face validity.

For construct validity, the validity is valid. The final semester test was made for testing listening and reading. But, due to the technical problem, the listening comprehension was not conducted by the students. To find the construct validity of the test, the test was analyzed by the concept of reading comprehension. Based on the classification of the final semester test, all reading items show a link to the traits of the reading test. This is the same as the content validity of the final semester test. The content validity of the final semester test is valid because all items in the reading comprehension are relevant to the syllabus in KTSP.

For face validity, it was evaluated by using the Guidelines for Constructing Multiple Choice Tests. So, if the test lacks face validity, it may not work as it should,

and may have to be redesigned. The results show that most of the items are not good and need to be revised.

In the output data of the ITEMAN, the result shows that the reliability coefficient of alpha is 0.448. Based on the criteria of the reliability of the test items, it is categorized as average/sufficient, that is, the test items whose alpha ranges from 0.401 – 0.700. It means that the test items in general if they are tested frequently under the same condition, they might result in similar outcome.

The test items are good if they are not too easy or not too difficult, or in average level. So, if the test is in the average level of difficulty, the test is good for the students. Related to the result of the level of difficulty in the output data of ITEMAN, some of the items fulfill the quality of a good item, but some do not.

Regarding with the item analysis using ITEMAN, it was found that the level of difficulty can be classified into four categories, that is, *good or directly usable*, *very difficult or needs revising, very easy or needs revising,* and *too difficult or needs dropping or total revision*. The criteria of the items which have the level of difficulty ranging from 0.300-0700 are categorized as *good or directly usable*. This class consists of 11 items (30%). There are eleven items that are *good*, that is 17, 18, 19, 21, 24, 29, 35, 38, 40, 41, 47. These items are recommended to be directly used without any prior revision. For the criteria *very difficult or needs revising*, the items have the level of difficulty ranging from 0.100-0.299. This class consists of 4 items (10%). There are four items that are very difficult, that is, 16, 26, 31, 49. These items need to be revised. As to the category *very easy or needs revising*, the items have the level of difficulty ranging from 0.701-0.900. This class consists of 6 items (20%).

There are seven items that are very easy, that is, 20, 21, 23, 25, 36, 45. These items also need to be revised. With reference to the criteria of the items which have the level of difficulty ranging from 0.000-0.099, the items are categorized as *too difficult or needs dropping or total revision*. This class consists of 14 items (40%). There are fourteen items that are *too difficult*, that is, 27, 28, 30, 32, 33, 34, 37, 39, 42, 43, 44, 46, 48, 50, therefore, they need dropping.

There are 6 items (17.1%) in the final semester test which have negative discrimination value, that is, 17, 19, 30, 31, 33, 38. It means that these items should be checked whether the key answer is correct. Related to the item analysis using ITEMAN, it was found that the test items whose discriminating power $\geq$ 0.400 is classified as *high*. There are 9 items (25.7%) that are *high*, that is, 23, 24, 25, 29, 35, 40, 41, 47, 49. These test items are recommended to be used as they can discriminate between the more knowledgeable from the less knowledgeable students. The criteria *average/without revising* is the items whose discriminating power ranges from 0.300-0.399. There are 2 items (5.7%) that do not need revising, that is, 16, 21. Concerning with the criteria *low/needs revising*, it points out that the items whose discriminating power ranges from 0.200-0.299. It was found that there are no items (0%) which involve in low discriminating power or need to be revised. The test items whose discriminating power range from 0.000-0.199 are categorized as *very low/needs dropping*. There are 18 items (51.5%) that are *too difficult*, that is, 18, 20, 22, 26, 27, 28, 32, 34, 36, 37, 39, 42, 43, 44, 45, 46, 48, 50.

Based on the results of the data analysis using ITEMAN, it was found that the alternative of the 35 items consisting of A, B, C, D, and E with the total of the

alternatives is 175, can be classified into three categories, that is, *very good, good enough or sufficient,* and *least/dropped, or needs revising.* With respect to the criteria *very good,* the alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.051-1.000. This class consists of 26 options (15%). These alternatives are recommended to be used without any prior revision. The alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.011-0.050 is categorized as *good enough or sufficient*. This class consists of 43 options (24.5%). These alternatives are recommended to be directly used, because they are chosen by at least 5% of the testees. Related to the criteria *least/dropped, or needs revising,* it is the alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.00-0.010. This class consists of 46 options (60.5%). These items should be revised before being tested.

5.2. Suggestions

In line with the conclusions above, some suggestions are proposed as follows:

1. Suggestions to the teachers
    a. According to the data gained, the teachers should be familiar with construct validity, content validity, and face validity in order that they can assess the quality of the test.
    b. The teacher should be good at the assessment from the aspects of material, construction, and language in order to improve the quality of the test.
    c. The teachers should be familiar with ITEMAN software program in order that they can assess the students' ability faster.

d. The teachers should be trained to use ITEMAN software program in order to improve the quality of the test.
e. The teachers should be familiar with all the terms related to the quality of the test items, such as, validity, reliability, prop. Correct (level of difficulty), point biserial (discriminating power), prop. Endorsing (options), distracters, key answers, alpha, and standard deviation.

2. Suggestions to other researchers
   a. It is suggested that the role of ITEMAN in determining the quality of multiple choice items is investigated further. It is also interesting to collect a larger or smaller data base for investigating whether there are more tendencies in determining the quality of items.
   b. Other researchers should replicate the current study in analyzing the quality of other test items, such as, Mid Semester Test, Final School Test (UAS), and National Examination (UN).