ANALYZING THE QUALITY OF THE FINAL SEMESTER TEST AT THE SECOND YEAR OF SMP NEGERI 2 SUMBEREJO IN 2016/2017 ACADEMIC YEAR

Nurul Amanah, Ujang Suparman, Ramlan Ginting

University of Lampung

Nurulamanah831@yahoo.com

Abstrak

Penelitian ini bertujuan untuk mengidentifikasi: validitas, keandalan, tingkat kesulitan, daya beda, kualitas pilihan di SMP Negeri 2 Sumberejo di tahun ajaran 2016/2017. Peneliti meneliti ujian akhir semester yang berjumlah 50 soal. Hasil penelitian menunjukan bahwa keandalan (alpha) adalah 0.592. Tingkat kesulitan terdiri dari 22 diterima, 21 item perlu diperbaiki, 7 item perlu dibuang. Daya beda terdiri dari 15 item diterima, 12 item perlu dibuang, dan 23 item perlu dicek. Kualitas pilihan adalah 26 pilihan diterima, 89 pilihan perlu diperbaiki dan perlu dibuang. Peneliti menyarankan agar soal-soal tersebut harus direvisi sesuai dengan yang disarankan oleh program.

The objectives of this research are to identify: the validity, the reliability, the level of difficulty, the discriminating power, the quality of the alternatives, of the final semester test at the second year of SMA Negeri 1 Purbolinggo in 2016/2017 academic year. The researcher investigated the final semester test consisting of 50 items. It was found the reliability is 0.592. The level of difficulty consists of 22 items acceptable, 21 items need revising, and 7 items need dropping. The discriminating power consists of 15 items acceptable, 12 items need revising, and 23 items need dropping. The quality of the alternatives is 26 options acceptable, 89 options need revising and need dropping. The researcher suggested that this questions should be revised as suggested by the program.

Keywords: *quality of the alternatives, reliability, validity.*

INTRODUCTION

Multiple choice testing is an efficient and effective way to assess a wide range ofknowledge, skills, attitudes, and abilities. By assessing the broader and morecomplexed competencies, multiple choice tests allow expansive and even deepcoverage of content in a relatively efficient way. The consideration behind thestatement is that it comes as the most standardized tests, including school or nationalexamination. Most profitable tests are mainly made up of multiple choice items.Besides, multiple choice tests may give a more accurate picture of how well studentshave met the standard.

As a matter of fact, the teachers sometimes added some additional competencies tomake the students easier to understand the subject. It means that the multiple choicetests might have a bad effect on overall curriculum and instruction. They stated that the multiple choice test was sometimes too easy and too difficult for the students.Because of that, what the students know in a subject was cut down from what themultiple choice test measured. The distracters in the multiple choice test might not beheterogeneous, which made the test weak. From the cases mentioned above, it can beconsidered that the multiple choice test might not discriminate the moreknowledgeable students from the less knowledgeable students.

The teachers are more expected to have an involvement in assessing the multiplechoice tests using the item analysis program as ITEMAN is considered useful. Inorder to utilize the program, the ITEMAN software program should be installed first.Another fact that motivated the writer to conduct this research was his own experience that proved his assessing multiple choice tests to have been easily analyzed by using the program. Because of that, the writer here put an effort on how to find some ways to utilize the program as a treatment to promote the assessment of multiple choice tests. Thus, this research was regarded to as a facilitative way for the teachers to analyze the final semester test.

The objectives of this research are to identify: the validity, the reliability, the level of difficulty, the discriminating power, the quality of the alternatives, of the final semester test at the second year of SMP Negeri in 2016/2017 academic year.

METHODS

This research used descriptive quantitative methods. The data were taken from the final semester test items created by MGMP in 2016/2017 academic year which consisted 50 items. The sample of this research was the students of VIII 3 at SMPN 2Sumberejo in 2016/2017 academic year. This class was taken by using *purposive sample*. The researcher needed to get a group of students who had the lowest score among others as the sample. The purpose was to determine the quality of the final semester test more accurately. Since the final semester test made by MGMP has been used for years by the school, it meant that the test was considered good. The researcher needed to know if the group of the lowest students had really bad scores due to the test, to the ability, or to the learning process. Due to that matter, the researcher chose VIII 3 as the sample of this research. The instrument was the final semester test; each item had four options

A., B., C., and D. There were 50 questions in the test. In order to know the quality of the final semester test, the researcher analyzed the test using traits of language skills and aspects of language, KTSP Curriculum, Guidelines for Constructing Multiple Choice Test, and ITEMAN software program.

RESULTS

This research was carried out in order to determine the quality of the final semester test identifying the validity, the reliability, the level of difficulty, the discrimination power, and the proportion of the alternatives. The final semester test was administered in class VII 3. The number of the students was 32 students. The final semester test was conducted on February 27th. There were 50 questions in the final semester test done by the students and analyzed by the researcher. To find out the construct validity of the test, the test was analyzed by the concept of reading comprehension.

Face validity is the tendency for a test to look like a test. The items based on the tendency for a test to look like a test is analyzed by using the guidelines for constructing multiple choice items. The face validity of the final semester test is categorized as not valid. Most of the items need to be revised, and some are good. But, according to Haladyna (2004:97), there are two categories of item whether the item correlates to the guidelines or not, that is, flawed and non-flawed items.

The output data of ITEMAN program shows the alpha (reliability of the test) is 0.592. With reference to 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.

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 is categorized as *good or directly usable*.This class consists of 22 items (44%). There are 16 items that are good, that is 4, 14, 17, 21, 22, 26, 27, 28, 31, 32, 33, 3, 37, 38, 44 and 46. For the criteria *very difficult or needs revising*, the items have the level of difficulty ranging from 0.900-0.100. This class consists of five items (10%). There are four items that are very difficult, that is, 16, 26, 31, 49. As to the category *very easy or needs revising*, the items have the level of difficulty ranging from 0.000-0.900. This class consists of 6 items (20%). There are seven items that are very easy, that is, 20, 21, 23, 25, 36, 45. 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 2 items (4%). There are two items that are too difficult, that is, 3 and 10.

There are 23 items (46%) in the final semester test which have negative discrimination value, that is, 1, 3, 4, 6, 7, 13, 14, 15, 21, 24, 25, 26, 27, 28, 33, 3, 37, 38, 45, 46, 48, 49 and 50. Related to theitem analysis using ITEMAN, it was found that the test items whose discriminating power ≥ 0.400 is classified as *high*. There are 9 items (25.7%) that are *high*, that is, 8, 9, 16, 17, 18, 19, 30, 35, 39, 40, 41, 42 and 43.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 (4%) that do not need revising, that is, 2 and 44. 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 12 items (24%) which involve in low discriminating power or need to be revised, that is item number 5, 10, 11, 12, 20, 22, 23, 29, 31, 32, 34 and 47. The test items whose discriminating power range from 0.000-0.199 are categorized as *very low/needs dropping*. There are 23 items (46%) that are *too difficult*, that is, 1, 3, 4, 6, 7, 13, 14, 15, 21, 24, 25, 26, 27, 28, 33, 3, 37, 38, 45, 46, 48, 49 and 50.

Based on the results of the data analysis using ITEMAN, it was found that the alternative of the 50 items consisting of A, B, C, and D with the total of the alternatives is 200, 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 10 options (15%). 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%). 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%). alternatives whose Prop. Endorsing (proportion of the answers) ranges from 0.00-0.010. This class consists of 46 options (60.5%).

DISCUSSION

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. As known that the test is considered valid if the test measures the object to be measured (Carmines and Zeller, 1979:17), it means that if the test items are good, the test has high validity. The construct validity and the content validity of the final semester test are valid, except face validity. Concerning with the previous research, the researchers did not analyze the test using construct validity, content validity, and face validity. Therefore, the similarities and differences from the previous research are not examined in this section.

For construct validity, the validity is valid. As stated by Brinberg & McGrath (1985:115), the term construct validity is used both for correspondence at the element level and at the relation level. To find the construct validity of the test, the test was analyzed by the concept of reading comprehension. According to O’neil (2009:23), a test is valid for anything with which it correlates. 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. According to O’Neill

(2009:26), face validity is a test looked like it would measure the desired ability or trait. 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 0.592. 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. Related to the previous research, Ratnaningsih (2009) gives the similar result with this research finding, that is, has good reliability. 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.

The findings of the research show that some of the items fulfill the criteria of the requirements of the quality of a good test item but some do not. With reference to the previous theories, Ariyana (2011) and Ratnaningsih (2009) presented the similar result with this research. They showed that more than 50% was good. But, Fitriyana (2013) gave different conclusion from this research. She had analyzed the multiple choice test resulting the test was deemed to be good enough since there were 15 items or 37.5 % of the good test. Negative discrimination would signal a possible key error (Haladyna, 2004:228). The result from the three previous theories did not elicit the key error, which means that there is negative value in the discriminating power. On the other hand, this research discovered that there were 6 items (17.1%) in the final semester test which had negative discrimination value, that is, 17, 19, 30, 31, 33, 38.

The number of the items which is considered to be under the category of sufficient/good enough covers 60.5%. Related to the previous theories, Ariyana (2011), Fitriyana (2013), and Ratnaningsih (2009) gave the similar result with this research. The three theories showed 82%, 67.5%, and 62% respectively. It indicated that the three theories had functional alternatives which were similar to this research.

CONCLUSIONS

Based on the results of the data analysis and discussions, some conclusions can be drawn as follows: the construct validity is valid, the content validity is valid, but the face validity is not valid. The reliability is 0.448. The level of difficulty consists of 11 items (30%) acceptable, 10 items (30%) need revising, and 14 items (40%) need dropping. The discriminating power consists of 11 items (31.4%) acceptable, 18 items (51.5%) need revising, and 6 items (17.1%) need dropping. The quality of the alternatives is 26 options (15%) acceptable, 89 options (85%) need revising and need dropping.

SUGGESTIONS

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

1. Suggestions to the teachers

   a. Concerning with the finding of the research that the the teachers should be familiar with and good at the assessment from the aspects of material, construction, and language in order to improve the quality of the test.

   b. The teachers should be familiar with and use ITEMAN software program in order to improve the quality of the test.

   c. 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. 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).

REFERENCES

Ariyana, L. T. 2011. *Analiiss butir soal ulangan akhir semester gasal ipa kelas ix smp di kabupaten grobogan*. Jurusan Biologi Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang.

Brinberg, D. & McGrath, J. E. 1985. *Validity and the research process*. beverly hills: sage publications.

Carmines, Edward G., & Richard A. Zeller. 1979. *Reliability and validity assessment*. beverly hills, ca: sage.

Fitriana, N. 2013. *Analisis kualitas butir soal ulangan akhir semester gasal mata pelajaran ipa kelas v mi sultan agung tahun pelajaran 2012/2013*. Universitas Islam Negeri Sunan Kalijaga, Yogyakarta.

Haladyna, T. M. 2004. *Developing and validating multiple-choice test items-3rd ed*. New Jersey: Lawrence Erlbaum Associates.

O'Neill, P. 2009. *A guide to college writing assessment.* Logan: Utah State University Press**.**

Ratnaningsih, D. J. 2009. *Analisis butir soal pilihan ganda ujian akhir semester mahasiswa di universitas terbuka dengan pendekatan teori tes klasik,* FMIPA-UT, Jl. Cabe Raya, Pondok Cabe, Pamulang, Kota Tangerang Selatan.