

AN ANALYSIS OF COLLEGE ENTRANCE TEST

By
Anni Fazlina *

Syiah Kuala University, Banda Aceh

ABSTRACT

This research is aimed at analyzing the quality of entrance test conducted by STKIP Bina Bangsa Getsempena year 2016-2017 in terms of validity, reliability, difficulty index, discrimination index, and the effectiveness of distractors. Content analysis was employed as research method in this study. Document analysis is a technique carried out to collect the data. The objects of this study were 25 multiple-choice items used in entrance test, with five options for each item. The data were analyzed manually. The result showed that English entrance test of STKIP Bina Bangsa Getsempena year 2016-2017 has a low predictive validity with correlation coefficient 0.044. Secondly, the *r-value* was 0.664, which is higher than *r-table* = 0.396. It is considered that the test has good reliability. Thirdly, the difficulty index was obtained that 4% of the test was easy with index 0.71-1.00, 64% was moderate with index 0.31-0.71, and 32% was difficult with index 0.00-0.31. Fourthly, most of items considered has perfect discrimination index, with the percentage is 84%, while there were also 4% with zero discriminaton and 12% with negative discrimination. Finally, 97% of distractors of STKIP Bina Bangsa Getsempena English entrance test were effective since they were chosen by more than 5% of the whole participants. There were 3% distractors that were considered ineffective as they were chosen by less than 5% of the whole participants. It means that the items in STKIP Bina Bangsa Getsempena English entrance test particularly in their difficulty index, discrimination index, and effectiveness of distractors are needed to be revised if the items will be used in next years' entrance test.

Keywords: *college entrance test, item analysis, undergraduate students*

* Corresponding author: daiyah88@gmail.com

INTRODUCTION

Education is a major factor for developing country. Through education, human resources can be developed. Improving the quality of education is the first step in the improvement of human resources. In this case, learning institutions such as school and college being a main priority to develop the quality of human resources. In the *Undang-Undang Republik Indonesia* (1989) states that the college is a continuation of secondary education that is organized to prepare students to become members of the community who have the academic ability and professionalism who can apply, develop and create science, technology and art.

Many colleges are now set up to compete against each other to produce the best graduates. However, the good quality of output can be initiated by obtaining the good quality of input. To get students with good quality, it should begin with the implementation of the selection of new admissions. Test arrangement, for example, needs to be reviewed in advance whether the establishment is in compliance with the rules for getting the best test to recruit good quality new students. Since the test will be taken by participants who have different levels of knowledge, then the test should represent the ability of all participants from easy to difficult level.

Teacher Training and Education College of Bina Bangsa Getsempena (STKIP BBG) is one of the private higher educations in Aceh. STKIP BBG established since September 5, 2003, based on the Decree of Directorate General of Higher Education No. 138/D/0/2003, signed by the Director General of Higher Education Satryo Soemantri Brojonegoro. There are six study programs available for prospective new students every year. The participants who wish to continue their studies in STKIP BBG come from various cities/districts in Aceh. The best students can be selected through a test selection of the new admissions. In general situation, testing is finding out how well something works, it will tell the level of knowledge or skill has been acquired. In this case, the test refers to a part of assessment used by STKIP Bina Bangsa Getsempena in selecting new competent students to attend the institution.

A test is supposed to measure someone's knowledge whether they already understand enough toward an issue. Therefore, testing is one of the powerful tools in measuring students' abilities as well as enhancing their attitudes towards knowledge. This notion is in line with Brown

(2004) who stated that a test is a method of measuring a person's ability, knowledge, or performance in a given domain. In the same line, Nitko (2004) defined that test is systematic procedure for observing and describing one or more characteristics of person with the aid of either a numerical or category system. Based on Cronbach in Azwar (2010) a test is a systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale or category system. It means that a test is a beneficial tool to evaluate the achievement and the ability of a person.

Test items constructed by the committee of new admission in multiple-choice form with five options for each question. In the preparation of the college admission test, STKIP Bina Bangsa Getsempena formed a committee that consists of some lecturers from each study programs. Testing is intended to know the applicants' mastery of knowledge, in the hope of prospective students have the basic concepts required to attend lectures at STKIP Bina Bangsa Getsempena in accordance with the chosen course of study.

The test is not only depend on the collection of test items but also determined by the measuring object, especially in view of the validity, reliability, level of difficulty, discrimination index and the effectiveness of distractors. A good quality test has some basic characteristics that can be used as standard, namely validity, reliability, difficulty index, discrimination and distractor. This opinion is in line with Cecil et al. (2009) who said that the reliability and validity of test scores are depended on the quality of the items of the test. It means that a test should be well constructed by considering its validity and reliability, while difficulty index, discrimination index, and distractor effectiveness. Those five characteristics will provide an accurate test toward test takers.

Based on the result of observation in entrance test for new students of STKIP Bina Bangsa Getsempena, the researcher found that the test given for new entrants has never been analyzed yet. It needs analyzing to find out the degree of quality of test items.

Research Questions

1. To what extent is the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017 valid?
2. What is the reliability index of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017?

3. What is the difficulty index of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017?
4. What is the discrimination index of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017?
5. How is the effectiveness of distractors of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017?

Research Objectives

1. To analyze to what extent the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017 valid.
2. To analyze the reliability index of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017.
3. To know the difficulty index of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017.
4. To know the discrimination index of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017.
5. To analyze the effectiveness of distractors of the entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017.

LITERATURE REVIEW

Concept of Test

Commonly, the test can be defined as a set of questions to be answered by test takers which is used as a tool by someone in measuring the test takers' ability or knowledge. It is also stated by Anderson et al. (1981) in the Encyclopedia of Educational Evaluation that test is a series of questions or exercises or other means of measuring the skill, knowledge, intelligence, capacities or aptitudes of an individual or group. The word "test" is derived from the Latin namely *testum* (Fresch & Wheaton, 2002; Singh, 1986). Broadly, test is defined by Brown (2004) as a method of measuring a person's ability, knowledge, or performance in a given domain. Furthermore, Cronbach (1980) also defined that test is systematic procedure for observing a persons' behaviour and describing it with the aid of a numerical scale or a category system.

Those definitions explain that a test is a tool with systematic procedure which consists of a group of questions or tasks to measure test takers' particular behavior by using a numerical scale or a particular category. The test is said to be a systematic procedure since the test must be compiled, executed and scored based on predetermined

rules. The test results will reveal specific behaviour of the test takers, then the information will be used as a basis for making an assessment. A test is also considered as a tool in which the test takers must answer all the questions in the test so that the results can be used to measure the knowledge or behavior of the test takers.

Test

In the educational environment, there are various types of tests which will possibly be chosen by educators. It really depends on the test makers' purpose and consideration. Generally, test can be classified into some types: based on the interpretation of scoring (norm-referenced test and criterion-referenced test); based on psychological aspect (intelligence test, aptitude test, personality test and achievement test); and based on the implementation (oral test, written test and performance test). Besides, there are two forms of the test commonly used by educators, (1) essay, and (2) objective.

Based on the interpretation of test scores, Cecil et al. (2009) states that score interpretation can be classified as either norm-reference or criterion-referenced; and this distinction refers to something to compare the examinees' performance. In the norm-referenced interpretation, the test takers result is compared to the other people (a reference group) in normative way. For example, in selecting five out of ten students who has good ability in listening skills, after having the test, all the students' result are compared each other and five of them who have the higher scores than others will be selected.

While criterion-referenced score interpretations, the examinees' performance is compared to a specified level of performance. For example, in National examination of Senior High School students, to pass the examination, a student must have the examination result at the equal or higher than standard score that has stated by the ministry without need to compare to the other students' result. In conclusion, norm-referenced interpretations are relative in giving the result for test takers, whereas criterion-referenced interpretations are absolute.

In psychological aspect, Yusrizal (2016) classifies test into four types; there are intelligence test, aptitude test, personality test, and achievement test. Intelligence test refers to measuring someones' ability in doing a task, since we know that intelligence allegedly depends on the level of ability gained from the environment in which someone grew up. Intelligence tests are usually used for measuring cognitive function. Similar to intelligence test, aptitude test also refers

to measuring someone ability but in certain part. This kind of test is used to measure specific abilities to learn or perform new tasks that may be relevant to learning or performance in a particular field. Therefore, the intelligence test is oriented toward the future, and this tests are commonly used for selection and placement processes.

Personality tests are also categorized in psychology because they are closely related to each other. This test aims to reveal the individual characteristics of the subject in measured aspects, such as how to think, feel or behave. Some personality tests are used to measure an individual's attitude, interest, even emotionality, or measure deviant behavior patterns. While achievement test, is a test used to evaluate what has been obtained in an activity. The achievement test relates to the purpose of teaching, which is what has been taught, as Gronlund and Robert (1995) have pointed out that achievement testing plays an important role in the school programs. This test is commonly used to see the extent to which students master the subject matter that has been taught, the quality or level achieved, the changes and progress of learning and the effectiveness of teaching strategies.

Based on the implementation, the test can be divided into three types, namely oral test, written test and performance test. Oral test is a term that is addressed to a situation where educators engage in a question and answer directly to learners. This test is commonly used to evaluate the ability of a learner in expressing ideas and knowledge orally. Written test - as the name suggested - is a test in which questions and answers generated in writing. The main focus of this test is having the written answers, even the question delivered by educators orally. This test is usually carried out in a room followed by a group of learners. While performance test is a test conducted in written form but the assignment is done by act or performance. This type of the test is usually used as a tool in the observations of behavior. The tool can be formed as observation sheet that is used to evaluate process or product of a practical activity.

Based on form, in academic assessment, the educators mostly use essay and multiple-choice form of test to measure the achievement and knowledge of learners. William and Stephan (1990) explained that essay test is an item format that requires the student to structure a rather long written response up to several paragraphs. It commonly contains more than one question in the test with requires completely long writing answers. The main purpose of using this test is to attempt the test takers' recalling the memory. By using this type of test, the test

makers can measure complex learning outcomes, emphasize integration and application of thinking and problem solving. Besides, essay test can be easily constructed, require less time for typing, duplicating or printing, can be written on board and also can be used as device for measuring and improving language and expression skills. In contrast, the examiners will be difficult to score objectively and will waste most of the time in examining the answer. Yusrizal (2016) agreed that the essay test is often referred as the subjective test since the process of assessment can affect the score of the test.

In contrast to essay test, the objective test has clear and unambiguous scoring criteria. Similar to its name, objective test can be objectively scored. There are two kinds of objective test: (a) using two alternatives, which is used in true/false items and completion type; and (b) using more than two alternatives which is used in matching items and multiple-choice questions. Multiple-choice test is one of the tests incorporated in objective test. It is one option of any available testing instruments that has been used in most standardized tests such as TOEFL, National Examination, and Employment Test since it can be set to a large number of participants. Multiple-choice test has its own characteristic. Hopkins and Antes (1979) explain that the multiple-choice item consists of a premise and a set of alternatives. The premise, known as the stem, is presented as a question or incomplete statement, which the student answers or completes by selecting one of several alternatives. The premise is usually followed by four or five alternatives. One of them is the correct answer and the rest of alternatives are as distractors. Cecil et al. (2009) mention that multiple choice items can be written as a direct question and incomplete sentence. An example of direct-question format is “Which river is the largest in the United States of America?”

Characteristics of a Good Test

Brown (1976) stated that a test must possess two characteristics if it is to be considered a good measuring instrument; the test must be reliable and valid. Yusrizal (2016) also mentioned that objectivity, practicability and economical as characteristics of a good test besides having validity and reliability. Validity and reliability are the major consideration in stating a good test. Besides, for multiple-choice form, the test makers also need to be concerned in item analysis. Cecil et al. (2009) stated that the reliability of test scores and the validity of the interpretation of test scores depend on the quality of the items on the

test, so the quality of an item of test can be judged by analyzing difficulty index, discrimination index and the effectiveness of distractors.

Validity

The word validity is derived from word “valid”. Gronlund (1995) as cited by Brown (2004) defined test validity as the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment. Whereas according to Hughes (2003), validity refers to whether the test measures accurately what it is intended to measure. In short, to be a valid, a test must provide consistently accurate measurement in order to be more appropriate, meaningful and useful assessment. It is also strengthen by Cohen et al. (2007), who stated that effective research is impossible or even “worthless” without the presence of validity.

To determine the categories of validity, it can be classified as suggested by Guilford (1956) in Yusrizal (2016) as follows:

0.80 < rxy 1.00 very high (very good)
0.60 < rxy 0.80 high (good)
0.40 < rxy 0.60 moderate (enough)
0.20 < rxy 0.40 low (less good)
0.00 < rxy 0.20 very low (bad)
rxy 0.00 not valid

Reliability

Reliability involves the consistency of test scores. Gronlund and Linn (1994) stated that reliability refers to consistency of measurement, that is how consistent test scores or other evaluation results are from one measurement to another. Furthermore, Deblassie (1974) mentioned that reliability is the extent to which a test measures consistently what it purposes to measure. The reliability of an item has a relation to validity. Hughes (2003) pointed out that if a test is not reliable, it cannot be valid. However, Alias (2005) argued that a test may be highly reliable but not necessarily valid, but a highly valid test is usually reliable.

Item Analysis

An analysis of test item is a systematic process to evaluate item test especially for objective test. It is needed to increase the quality of a test. A good quality of test will be obtained based on empirical data on the items concerned. The analyses of item tests are: the difficulty level, the discrimination index and the effectiveness of distractors.

Difficulty Level

A good test item should have a certain degree of difficulty. Index of difficulty is numerical figures that differentiate between the students who answer correctly and those who answer the question incorrectly. Heaton (1989) states that the index of difficulty (facility value) of an item simply shows how easy or difficult the particular item provided in the test. It is generally expressed by the fraction of students who can answer correctly. It means that the students who can answer the question correctly play the key role in determining the index of difficulty of test items.

A good test item has a certain degree of difficulty. It should not be too easy nor too difficult. An expected degree of a test item is to be in moderate level of difficulty index. As Heaton suggests the difficulty 0.00 – 0.30 being difficult level, 0.31 - 0.70 being moderate level and 0.71 – 1.00 being easy level. If an item is too easy or too difficult, then the score distribution will be hard to identify, so that the reliability of the test is being questionable. Besides, Bart (2011) supports that it is also difficult to differentiate between those who have done well and those who have done poorly in the test. The difficulty level analysis of test items is needed to check whether the test is too easy or too difficult to answer so that the teacher can revise or eliminate the test item.

Discrimination Index

Ahmann and Glock (1967) defined discrimination as an index that shows its ability to differentiate between pupils who achieved well (the upper group) and those who have achieved poorly (the lower group). The tests that do not have discriminating power will not yield the proper description of the student's ability. The test which does not have the discriminating power will not give a reliable interpretation of the students' real ability. The discrimination index range is between -1 and +1. The closer the index to +1, the more effectively the item distinguishes between the two groups of students. Arthur (2013) argue that sometimes an item will discriminate negatively, which means it

discriminates in the wrong direction by interpreting the successful students do poorly in the test and the poor students do well in the test. Such an item should be revised and eliminated from scoring as it indicates that the lower group students actually selected the correct answer more frequently than the upper group of students.

Distractor

Each multiple choice test item provides alternatives that have to be chosen by pupils to answer. One option is a key answer and others are distractors. According to Fulcher and Davidson (2007), distractor analysis is another useful step in reviewing the effectiveness of a test item. If the distractor does not work, it will lead them to select true answer. To know whether distractor works, it can be seen on how many students fail to choose true answer. A distractor will be acceptable if it should be attract at least five percent of all participants. If no one selects a distractor, it is important to revise the alternative.

Research Design

Descriptive analysis is the method used in this research. Sudjana (2004) defined that descriptive method is a study that aims to describe or explain the events ongoing at the time of the study, regardless of before and after. Furthermore, Sukardi (2003) stated that descriptive research is generally done with the aim to describe systematically the facts and characteristics of the object or subject under study accurately. Moreover, this research also used quantitative approach to answer the research question that has been determined prior to the study that is to investigate the university entrance test items whether they meet the criteria of a good test; validity, reliability, difficulty index, discrimination index and distractor. Statistic was employed in this study to find the result of study.

Object of the study

The object of this research was the English item tests given for entrance test taken by new entrants of STKIP Bina Bangsa Getsempena year 2016-2017. The entrance test consists of 25 multiple choice items with five alternatives for each item.

Research instrument

Research instrument refers to the selected aids employed by researcher in collecting data. It denotes such a main object in

conducting the study as much as acquiring relevant or irrelevant data depends on the instrument. Concerning with research instrument, in this study a document was used to support and to get a result of research. The instrument would be used to answer all research questions.

There were three kinds of documents used in this study, they are (a) the English entrance test items, (b) students' odd semester summative test score, and (c) the applicants' answer sheets.

Data Collection Procedure

There are several steps taken during the completion of data collection. First, the students' answer sheets of the English entrance test of STKIP Bina Bangsa Getsempena were collected. Next, the answer keys which were given by the committee were checked to assure that it is error-free because these answer keys were the reference when checking the students' answer sheets.

Later on, in determining the validity, the final score of the students' first semester on Structure test were analyzed to find the degree of predictive validity. The data from students' answer sheets were analyzed to estimate the degree of reliability, difficulty index, discrimination index and the effectiveness of distractors. In short, all would be analyzed to find out the quality of the STKIP Bina Bangsa Getsempena entrance test. Finally, each result of analysis would be reported systematically.

Data Analysis

As mentioned before, to answer the first research question concerning validity of the test, only predictive validity would be the focus. The test scores were correlated to the final score of students' first semester result. It would be a proof whether the test constructed by the committee had predictability to the success of new students of STKIP Bina Bangsa Getsempena. Experts or supervisors were involved to help the researcher in analyzing test validity. The *Pearson Product Moment* was used to find out the correlation coefficient of the test. The criteria for validity used in this research is as in the following table:

Table 1. Validity Criteria suggested by Guilford (1956)

| Achievement | Criteria |
|------------------------|-----------|
| $0.80 < r_{xy} < 1.00$ | Very Good |
| $0.60 < r_{xy} < 0.80$ | Good |

| | |
|-----------------------------|-----------|
| 0.40 < r _{xy} 0.60 | Enough |
| 0.20 < r _{xy} 0.40 | Less Good |
| 0.00 < r _{xy} 0.20 | Bad |
| r _{xy} 0.00 | Not Valid |

Furthermore, identifying the reliability of test items was done by using *Kuder Richardson Formula 21* or commonly known as *KR21 Formula* as follows:

$$r_{21} = \frac{n}{n-1} \left(1 - \frac{X(n-X)}{nS_t^2} \right)$$

where:

r_{21} = the reliability of test

n = number of questions

x = the average test score

S = total score variant

The *r* (correlation coefficient) refers to critical value shown in *r table* (please see Appendix 5) with significant level 95%. For the number of subject N= 25, the critical index should be equal to or higher than 0.396 for the test items to be considered as reliable.

The following analysis is concerning with item difficulty. In order to examine the difficulty level of the test, the formula from Heaton (1989) was used:

$$FV = \frac{R}{N}$$

where:

FV = facility value or index of difficulty

R = number of correct answer

N = number of test takers

Heaton (1989) also suggests the following difficulty indices:

Table 2. Difficulty Index

| | |
|-----------|-------------|
| Difficult | 0.00 – 0.30 |
| Moderate | 0.31 – 0.70 |
| Easy | 0.71 – 1. |

In determining the discrimination power, the researcher used a formula which is also suggested by Heaton (1989) was used as follows:

$$D = \frac{\text{Correct } U - \text{Correct } L}{n}$$

where:

D = discrimination index

U = upper half

L = lower half

n = number of testees either in U or L or $(\frac{1}{2}N)$

Heaton (1989) suggests the following discrimination indices:

Table 3. Discrimination index

| | |
|----|---|
| +1 | The item discriminates perfectly |
| 0 | The item does not discriminate at all |
| -1 | The item discriminates in the wrong way |

Finally, in answering the fifth research question which is about scrutinizing the effectiveness of distractors, the researcher provides Sudijono's (2011) exemplification on tabulating the dispersion of the distractors below:

Table 4. Exemplification of Distractor Dispersion

| Item Number | Alternatives | | | | Note |
|-------------|--------------|----------|----------|----------|---------------------|
| | Option A | Option B | Option C | Option D | |
| 1 | 4 | 6 | 5 | (30) | () = correct answer |
| 2 | 1 | (44) | 2 | 1 | |
| 3 | 1 | 1 | (10) | 37 | |
| ... | ... | ... | ... | ... | |

Indeed, there must be at least 5% of the whole testees choose the distractor, so the claim that the distractor is good to be addressed. Otherwise, the distractors must be revised or dropped. In establishing

the percentage, the researcher will use the percentage formula as suggested by Sugiono (2007):

$$P = \frac{F}{N} \times 100\%$$

where:

P = numeral percentage

F = the frequency

N = number of testees

RESULTS AND DISCUSSIONS

There was one kind of test validity analyzed, namely predictive validity. Predictive validity refers to a measurement of how well a test predicts abilities of respondents. For this research, predictive validity is used to see how valid the test result to predict the students' scores in the future. As the English Entrance test's items of STKIP Bina Bangsa Getsempena year 2016-2017 encompass structure/grammar skill, therefore, to analyze the predictive validity, the students' structure final score of the first semester was needed to find out the predictive validity of the test.

To see the predictive validity percentage from the test, the correlation between the entrance test score and the student score after first semester (*see Appendix 3*), was calculated as follows:

$$R_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$$

$$R_{xy} = \frac{42.104440 - (3321)(1316)}{\sqrt{(42.277221 - (3321)^2)(42.49808 - (1316)^2)}}$$

$$R_{xy} = \frac{4386480 - 4370436}{\sqrt{(11643282 - (11029041)(2091936 - (1731856)}}$$

$$R_{xy} = \frac{16044}{\sqrt{(614241)(360080)}}$$

$$R_{xy} = \frac{16044}{\sqrt{131175898280}}$$

$$R_{xy} = \frac{16044}{362182}$$

$$R_{xy} = 0.044$$

Seeing the fact that the achievement coefficient validity falls between $0.00 < 0.20$. By referring to the criteria stated in Chapter III – suggested by Guilford (1956) - it means that the predictive validity of STKIP Bina Bangsa Getsempena is in bad category. Therefore, the

English Entrance Test of STKIP Bina Bangsa Getsempena has a low predictive validity.

Reliability of the Test

The next test result presented is the reliability index of the English Entrance test for new admissions of STKIP Bina Bangsa Getsempena year 2016-2017. In determining the test reliability, first researcher listed correct answer from each student for each item. Later, after she computed the result into KR-21 formula, she found that the test items are reliable. The computation is as in the following:

$$r_{21} = \frac{n}{n-1} \left(1 - \frac{X(n-X)}{nS_t^2}\right)$$
$$r_{21} = \frac{48}{48-1} \left(1 - \frac{8.9(48-8.9)}{48(20.76)}\right)$$
$$r_{21} = 1.021 (1 - 0.349)$$
$$r_{21} = 0.664 (\geq 0.396)$$

Based on computation, the *r-table* for subject 25 (since there were 25 question items) at the level of significant 95% is 0.396. It means the test items are considered reliable for it has correlation coefficient 0.664, which is clearly higher than *r-table* value 0.396.

Difficulty Index of The Test

By analyzing the difficulty index of the English Entrance test, the test makers could identify which items of the English Entrance test items for new admissions of STKIP Bina Bangsa Getsempena year 2016-2017 are categorized as easy, moderate, and difficult items. To obtain the index of difficulty from each question item, the researcher divided the score into two groups, upper group and lower group. Upper group is the group with higher scores, meanwhile lower group is the group with lower scores. Then, the correct answer from both groups were calculated and classified. As mention earlier in Chapter 3, the question item is considered difficult if its index is between 0.00-0.30; it is considered moderate - which means not too difficult nor too easy- is its index is 0.31-0.70; and lastly, the item considered as easy if it has the index of 0.71-1.00.

From the data analysis, the difficulty level of each of the question item can be provided in the following table:

Table 5. Difficulty Index of the Test

| Index of Difficulty | Interpretation | Total | Items |
|----------------------------|-----------------------|--------------|---|
| 0.00 – 0.31 | Difficult | 8 | 4, 5, 15, 16, 19, 22, 23, 25 |
| 0.31 – 0.70 | Moderate | 16 | 1, 2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 17, 18, 20, 21, 24 |
| 0.71 – 1.00 | Easy | 1 | 10 |

The data in the table shows that there is one item categorized easy, namely item number 10, since most of the new admissions were able to answer it correctly. The index difficulty of question number 10 is 0.72. Thirty five out of forty eight new admissions answered it correctly and only 13 new admissions did it incorrectly.

Next, eight test items belong to difficult item, namely items number 4, 5, 15, 16, 19, 22, 23 and 25. Question number 4 was answered correctly only by 14 new admissions and the rest answered it wrong, with the index of difficulty is 0.29. Question number 5 was answered correctly only by 12 new admissions and majority of the new admissions answered it wrong, with the index of difficulty is 0.25. Question number 15 was answered correctly by 13 new admissions and 35 of them answered it wrong, with the index of difficulty is 0.27. Question number 16 was answered correctly only by 7 new admissions and 41 new admissions answered it wrong, with the index of difficulty is 0.14. Question number 19 was answered correctly by 14 new admissions and 34 new admissions answered it wrong, with the index of difficulty is 0.29. Question number 22 was answered correctly by 11 new admissions and 37 new admissions answered it wrong, with the index of difficulty is 0.22. Question number 23 was answered correctly by 14 new admissions and 34 new admissions answered it wrong, with the index of difficulty is 0.29. Question number 25 was answered correctly by 9 new admissions and 39 new admissions answered it wrong, with the index of difficulty is 0.18.

Finally, there are 16 questions belong to moderate level. They are items number 1, 2, 3, 6, 7, 8, 9, 11, 12, 13, 14, 18, 17, 20, 21, and 24. Question number 1, 2, and 6 were answered correctly by 16 new admissions and 32 of them answered it wrong, with the index of difficulty is 0.33. Question number 3 and 18 were answered correctly by 17 new admissions and 31 new admissions answered it wrong, with the index of difficulty is 0.35. Question number 7 and 24 were

answered correctly by 21 new admissions and the rest answered it wrong, with the index of difficulty is 0.43. Question number 8 was answered correctly by 25 new admissions and 23 new admissions answered it wrong, with the index of difficulty is 0.52. Question number 9 and 11 were answered correctly by 23 new admissions and 25 new admissions answered it wrong, with the index of difficulty is 0.47. Question number 12 and 14 were answered correctly by 20 new admissions and 28 new admissions answered it wrong, with the index of difficulty is 0.41. Question number 13 and 17 were answered correctly by 15 new admissions and majority of the new admissions answered it wrong, with the index of difficulty is 0.31. Question number 20 was answered correctly by 18 new admissions and 30 new admissions answered it wrong, with the index of difficulty is 0.37. Question number 21 was answered correctly by 19 new admissions and 29 new admissions answered it wrong, with the index of difficulty is 0.39. Question number 24 was answered correctly by 21 new admissions and 27 new admissions answered it wrong, with the index of difficulty is 0.43. Looking at the number of moderate test items in the table is larger than others, it means that the difficulty index of English Entrance test items designed by the committee of STKIP Bina Bangsa Getsempena is categorized as moderate level.

Discrimination Index of the Test

Discrimination index is estimated to know whether the English Entrance test for new admissions of STKIP Bina Bangsa Getsempena can differentiate both the upper group and lower group. To analyze the discrimination index, the total new admissions are separated into two groups – i.e. upper group and lower group. The result of discrimination index analysis is presented in Table 6.

Table 6. Discrimination Index of the test

| Index of Discrimination | Interpretation | Total | Items |
|--------------------------------|-----------------------|--------------|--|
| +1 | Perfect | 21 | 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 21, 23, 24, 25 |
| 0 | No | 1 | 20 |
| -1 | Wrong | 3 | 6, 15, 22 |

It can be seen from the data above that there is one item test that has no any discrimination between upper and lower group of new admissions, namely item number 20. Besides, there are three item tests that have wrong discrimination index, namely item number 6, 15 and 22. However, the rest 21 items of 25 item tests consider to have perfect discrimination index between both upper and lower group of all new admissions of English Entrance Test of STKIP Bina Bangsa Getsempena year 2016-2017.

Specifically, question 20 is considered to have no discrimination since it was answered correctly by 9 new admissions from upper group and also 9 new admissions from the lower group.

The three questions that are considered to have wrong discrimination are questions number 6, 15 and 22. Question 6 was answered correctly by 6 new admissions from upper group and 10 new admissions from the lower group. Question number 15 was answered correctly by 6 new admissions from upper group and 7 new admissions from the lower group. Question number 22 was answered correctly by 5 new admissions from upper group and 6 new admissions from the lower group.

The rest 21 questions considered to have perfect discrimination; they are: question number 1 was answered correctly by 11 new admissions from upper group and 5 new admissions from the lower group, question number 2 was answered correctly by 10 new admissions from upper group and 6 new admissions from the lower group, question number 3 was answered correctly by 12 new admissions from upper group and 5 new admissions from the lower group, question number 4 was answered correctly by 10 new admissions from upper group and 4 new admissions from the lower group, question number 5 was answered correctly by 9 new admissions from upper group and 3 new admissions from the lower group, question number 7 was answered correctly by 12 new admissions from upper group and 9 new admissions from the lower group, question number 8 was answered correctly by 15 new admissions from upper group and 10 new admissions from the lower group.

Then, question number 9 was answered correctly by 12 new admissions from upper group and 11 new admissions from the lower group, question number 10 was answered correctly by 20 new admissions from upper group and 15 new admissions from the lower group, question number 11 was answered correctly by 15 new admissions from upper group and 8 new admissions from the lower

group, question number 12 was answered correctly by 16 new admissions from upper group and 4 new admissions from the lower group, question number 13 was answered correctly by 8 new admissions from upper group and 7 new admissions from the lower group, question number 14 was answered correctly by 14 new admissions from upper group and 6 new admissions from the lower group, question number 16 was answered correctly by 4 new admissions from upper group and 3 new admissions from the lower group, question number 17 was answered correctly by 10 new admissions from upper group and 5 new admissions from the lower group, question number 18 was answered correctly by 12 new admissions from upper group and 5 new admissions from the lower group.

Next, question number 19 was answered correctly by 8 new admissions from upper group and 6 new admissions from the lower group, question number 21 was answered correctly by 14 new admissions from upper group and 5 new admissions from the lower group, question number 23 was answered correctly by 9 new admissions from upper group and 5 new admissions from the lower group, question number 24 was answered correctly by 14 new admissions from upper group and 7 new admissions from the lower group; and question number 25 was answered correctly by 5 new admissions from upper group and 4 new admissions from the lower group. For clearer information.

Effectiveness of Distractors of the Test

Distractors in multiple choices are the incorrect options that distract the test takers from the correct answer. A good distractor will attract more test takers from the lower group than upper group. The STKIP Bina Bangsa Getsempena Entrance Test year 2016-2017 has 25 multiple-choice items. Each item has five options or alternatives, they are option a, option b, option c, option d, and option e.

From the analysis, it is found that from one hundred distractors, since there are twenty-five multiple-choice items with five options for each item, four options as distractors, and one option as key answer, only 3% of them are not effective. The distractors that are not effective are found in question 10, question 13, and question 22.

Based on the result analysis, it shows that all distractors for English subject of STKIP Bina Bangsa Getsempena entrance test are plausible since most of the distractors are able to attract new admission

from the right answer. In conclusion, based on the result of data analysis, the effectiveness of distractors of English Entrance test for new admissions of STKIP Bina Bangsa Getsempena can be concluded as follows:

Table 7. The Effectiveness of Distractors of the Test

| Category | Effective | Ineffective | Total Distractors |
|----------------|-----------|-------------|-------------------|
| Percentage (%) | 97 | 3 | 100 |

CONCLUSIONS AND SUGGESTIONS

Conclusions

Based on the results of data analysis, the researcher draws five following conclusions related to the research questions can be drawn:

1. The validity of English Entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017 is low. The result indicates that the predictive validity is in very low category (0.044 or < 0.20).
2. The reliability index of English Entrance test of STKIP Bina Bangsa Getsempena academic year 2016-2017 meets the standard/ criteria of reliability index, since it has correlation coefficient 0.664 which is higher than r-table value which is 0.396 for subject of 25 and significant level at 95 percent. It means that the test is up to standard because it has a consistent degree of reliability.
3. The difficulty index of STKIP Bina Bangsa Getsempena academic year 2016-2017 is categorized in moderate level since 64 percent or 16 of test items belong to moderate level. Only 4 percent or 1 item that belongs to easy level and 32 percent or 8 items that belongs to difficult level.
4. The discrimination index of STKIP Bina Bangsa Getsempena academic year 2016-2017 has good discriminating index, there is 84 percent or 21 questions with perfect discrimination index, 0 percent or 0 question with no discrimination index and 12 percent or 3 question with wrong discrimination index.
5. The effectiveness of distractors of STKIP Bina Bangsa Getsempena academic year 2016-2017 is good. There are 97 percent of the whole distractors that are effective and were

chosen by more than five percent of the test takers. And only 3 percent of the whole distractors are not effective since it was chosen by less than five percent of the test takers.

Finally, to conclude all results, most of the items in English Entrance test of STKIP Bina Bangsa Getsempera academic year 2016-2017 can be used as it was; however, some of the items are can be used after they are revised, particularly related to their index of difficulty, discrimination index, and the effectiveness of distractors.

Suggestions

To follow up the research findings, there are some suggestions to address to the committee of STKIP Bina Bangsa Getsempera academic year 2016-2017, teachers, lecturers, and other researchers who are interested in the field of English test evaluation. The suggestions are as in the following:

a. To the committee

The committee who design the test should check item analysis of the questions they administer to see the ability of new admissions. This is intended to see the test appropriateness in assessing new admissions' ability in the English language, especially in Structure part.

b. To chairman of the college

By having the committee who designed the test, it does not mean that the chairman of the college can let everything be processed by the committee, but should encourage them to check the level of difficulty, discrimination index, and effectiveness of distractors before the questions are distributed to new admissions. It is useful to know that this test can determine the level of understanding of prospective students.

c. To the other researchers

The researcher hope there will be other researchers who will examine the evaluation of a test, even continue this research. Because the evaluation of the test is a vital thing in order to have the appropriate test that is expected.

All test makers have to know how to analyze the item test to improve the quality of the test itself. Besides, the English Entrance test should be tried out to make sure the test meet a good test criterion. The test makers should provide neither too easy nor too difficulty item, so that the item will meet an ideal difficulty index. Besides, the item must be able to differentiate between upper group and lower group of test takers by having the discrimination index near to +1. Finally, the

conformity in the writing of stems and distractors is also very necessary to be considered in order to avoid errors in the writing according to standards set by experts and various common mistakes can be avoided.

REFERENCES

- Aiken, Lewis R. (1994). *Psychological testing an assessment (8th Edition)*. Boston: Allyn and Bacon.
- Alias, M. (2005). *Assessment of learning outcomes: Validity and reliability of classroom tests. A Journal from World Transactions on Engineering and Technology Education*, 4(2), 235-238.
- Anderson, S. B. (1981). *Encyclopedia of education evaluation*. San Fransisco: Jossey-Bass, Inc Publisher.
- Arikunto, S. (2009). *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara.
- Arthur, J. (2013). The history of psychological testings. *Behaviour Research Methods*, 2(17), 229-231.
- Azwar, S. (2010). *Tes prestasi: Fungsi dan pengembangan pengukuran prestasi belajar* (2nd edition). Pustaka Pelajar.
- Azwar, S. (2012). *Reliabilitas dan validitas*. Yogyakarta: Pustaka Pelajar.
- Bachman, L. F. (1990). *Fundamental consideration in language testing*. Oxford: Oxford University Press.
- Badan Akreditasi Nasional Perguruan Tinggi. *Direktori Hasil Akreditasi Program Studi*. Retrieved March 15, 2017 from https://banpt.or.id/direktori/prodi/pencarian_prodi.
- Bart, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree, and rule based predictive models. *Decision Support System*, 51(3), 141-154.
- Brown, D. H. (2004). *Language assessment: Principles and classroom practice*. New York: Longman.
- Brown, F. G. (1976). *Principles of educational and psychological testing (2nd Edition)*. New York: Holt, Rinehart and Winston
- Bryman, A. (2001). *Social research methods*. New York: Oxford University Press Inc.
- Cecil R., Reynolds, Ronald B. L., & Victor, W. (2009). *Measurement and assesment in education*. New Jersey: Pearson Education. Inc.
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research method in education*. London: Routledge Taylor and Francis Group.

- Cronbach, L. J. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Fransisco: Jossey-Bass.
- Deblassie, R. R. (1974). *Measuring and evaluating pupil progress*. New York: MSS Information Corporation.
- DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, 2(2), 1-23.
- Djaali, & Mulyono, P. (2007). *Pengukuran dalam Bidang Pendidikan*. Jakarta: Grasindo.
- Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *An E-article of PARE (Practical Assessment, Research and Evaluation)*, 18(3).
- Fresch, Mary Jo & Aileen Wheaton. 2002. *Teaching and Assessing Spelling*. Ohio. Scholastic Inc.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London: Routledge.
- Gronlund, N. E., & Linn, R. L. (1995). *Measurement and assessment in teaching (Seventh Edition)*. Ohio: Merrill, an imprint of Prentice Hall.
- Heaton, J. B. (1977). *English language tests*. Singapore: Longman.
- Heaton, J. B. (1989). *Writing English language tests*. New York: Longman.
- Hill, C., & Parry. K. (2014). *From testing to assessment: English as international language*. London: Routledge.
- Hopkins, C. D., & Antes, R. L. (1979). *Classroom testing: Construction*. Itasca, Illinois 60143. F.E: Peacock Publisher, Inc.
- Hughes, A. (2003). *Testing for language teacher*. Cambridge: Cambridge University Press.
- Ibrahim, M. (2013). Kualitas Tes Ujian Nasional Matematika Siswa Smp Negeri Di Kabupaten Buton Utara Tahun Ajaran 2011/2012. *Jurnal Penelitian Pendidikan Matematika*, 1(1), 1-10.
- Joshua, F. A. (2003). *Introduction to test construction in the social and behavioural sciences*. Rowman and Littlefield.
- Key, J. P. (1997). *Descriptive research*. Retrieved April 29, 2014, from <http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage110.htm>.
- Khalifa, H., & Weir, C. J. (2009). *Studies in language testing*. Cambridge: Cambridge University Press.

- McBeath, R. J. (Editor). (1992). *Instructing and evaluating in higher education*. New Jersey: Educational Technology Publications.
- Nitko, A. J. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Pearson.
- Ramadhana, M. A. (2011). *An analysis of the english test used in final examination at SMA Pesantren IMMIM Makassar*. Unpublished thesis.
- Rusma, S. (2015). *The analysis of English summative test at MAN Boarding School Meulaboh I*. Unpublished Thesis.
- Singh, A. K. (1986). *Tests, measurement and research methods in behavioral sciences*. New Delhi: Tata McGraw Hill.
- Sudijono. (2006). *Pengantar evaluasi pendidikan*. Jakarta: PT Raja Grafindo Persada.
- Sudjana, N. (2004). *Penilaian hasil proses belajar mengajar*. Bandung: Remaja Rosdakarya.
- Sukardi. (2003). *Metodologi penelitian pendidikan: Kompetensi dan praktiknya*. Jakarta: PT Bumi Aksara.
- Surgenor, P. (2010). *Teaching toolkit*. Retrieved Thursday 8, 2015 from <http://www.ucd.ie/t4cms/UCDTLT0029.pdf>.
- Suryabrata, S. (2000). *Pengembangan alat ukur psikologis*. Yogyakarta: Penerbit Andi.
- Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in Psychology and education*. New York: John Willey & Sons.
- Undang-undang Republik Indonesia, Nomor 2 Tahun 1989 tentang *Sistem Pendidikan Nasional, Pasal 16*. Retrieved February 24, 2017 from <http://www.dpr.go.id/dokjdih/document/uu/591.pdf>.
- Yusrizal. (2016). *Pengukuran & evaluasi hasil dan proses belajar*. Yogyakarta: Pale Media Prima.
- Wainer, H. I. B. (2013). *Test Validity*. Routledge.