**Analyzing Validity and Reliability Instruments of Teacher Made Test for English Subject**

**Nida Mahdiana Zahiroh**1, **Imam Ghozali**2, **Moh.Arif Mahbub**3

[zahirohnida@gmail.com](mailto:zahirohnida@gmail.com)

**Abstract** The development of education world not only supported by the factors of teaching such as teaching methods, learning process and others but also influenced by the factors which doesn’t related with the teaching process at the class that is evaluation of learning. The writer conducted observation in Eight grade at MTs Bustanul Ulum Kasiyan Timur. The objectives of this research are to describe the characteristics of the item of teacher made test for English subject for Eight grade in MTs Bustanul Ulum Kasiyan Timur – Puger in the 2014/2015 Academic Year on daily test 1, daily test 2 and daily test 3 which are include validity and reliability. Type of this research is descriptive research with analyzing the item of question, reviewed from its content validity based on the syllabus and reviewed from theirs face and construct validity based on the question made by teacher an reviewed its criterion-related validity based on the answer of the students at MTs Bustanul Ulum. Based on the data analysis which is done, so it can be concluded that on generally the item of the teacher made test on daily test 1, 2 and 3 has not fulfill the criteria of content validity. From face validity on daily tests 1, 2 and 3 the teacher using layout well. On construct validity, the teacher using different skill to make the question between daily test 1, 2 and 3. And from Criterion-related validity on daily tests 1, 2 and 3 has not fulfill standard of KKM. And for reliability has not reliable because on daily test 2 and 3 have low reliability.

**Keywords:** Teacher Made Test, Validity, Reliability

**Introduction**

English is a tool of communication to get information and it can be used in formal education as academic subject matter. In the global era, English is increasingly needed because it is one of the International languages mostly used in the world. English language become the first foreign language that should be taught to the English student in every level of education in Indonesia. English is taught as compulsory subject in elementary, junior and senior high schools, and as a complementary subject in university. The purpose of teaching English in Indonesia is to develop the communication skills especially in speaking, writing, listening, and reading, teacher applying evaluation to measure how far the students undertand about the material.

The purpose of evaluation is to get accurately information about students’ ability, so the teacher can try to do the next act. Evaluation of education not only to measure but also to get the value, but to get a score of value we need to do the next evaluation. From global definition of the measuring values, can be help to clarify the instructional, determining of the student’s needs, and determining a success of the students in the learning process. A teacher can understand how far the student can understand about the material by assessing their students. So, with this activity, the teacher is hoped to know about the students’ difficulties about the material. One activity that the teacher can do is give a test. In the teacher made test, the teacher, who make the test should know and master the principles and the steps that must be applied in making the test. The teacher made test was used to measure her student's achievement on the objectives give after finishing the teacher learning progress. Therefore, she also states that the teacher-made test has average or lower reliability than standardized test According to the background above, this research must do because the teachers seldom analyze the quality of evaluation scoring such as validity and reliability. This research will be doing in MTs Bustanul Ulum KasiyanTimur – Puger. Based on the observation when the Teaching Practice since 22nd of September until 22nd of November 2014 was done in MTs Bustanul Ulum Kasiyan Timur – Puger, the teacher almost never analyzes the quality of evaluation instruments. So, the writer is interested to carry out the research entitled “Analyzing Validity and Reliability instrument of the Teacher Made Test for English Subject for Eightgrade in MTs Bustanul Ulum Kasiyan Timur – Puger.

**Research Problem**

* + 1. To what extent the quality of the English teacher-made test with respect to their validity?

1. To what extent the quality of the English teacher-made test with respect to their reliability?

**Research Objectives**

* + 1. To describe or present the quality of the English teacher-made test with respect to their validity;

1. To describe or present the quality of the English teacher-made test with respect to their reliability;

**Literature Review**

(Brown, 2003:3) explain that “a test in simple terms is a method of measuring a person’s ability, knowledge or performance in a given domain”. The following lines will discuss about the types of language test proposed by Harmer (2001:321)

**Types of Test**

1. Placement Tests

Usually based on syllabuses and materials the students will follow and use once their level has been decided on, these test grammar and vocabulary knowledge and assess students’ productive and receptive skills.

1. Diagnostic Test

Diagnostic skills can be used to expose learners’ difficulties, gaps in their knowledge, and skill deficiencies during a course. Thus, when we know what the problems are, we can do something about them.

1. Progress or Achievement Test

These tests are designed to measure learners’ language and skill progress in relation to the syllabus they have been following. Achievement tests only work if they contain item types which the students are familiar with. This doesn’t mean that in reading test, for example, we give them texts they have been seen before, but it does mean providing them with similar texts and familiar task types. Achievement tests at the end of a term (like progress tests at the end of a unit, a fortnight, etc.) should reflect progress, not failure.

1. Proficiency Test

Proficiency tests give a general picture of a student’s knowledge and ability (rather than measure progress). They are frequently used as stages people have to reach if they want to be admitted to a foreign university, get a job, or obtain some kind of certificate.

**Skills in the Teacher Made Test**

1. Writing test

Written test is the one of the basic skills of the English Language. It is generally considered one of the most difficult that other skills for foreign language student. The description of the Certificate in Writing (Level 1) may not include the complete set of spesifications for that test but it shows what specifications for a writing test may look like. (Hughes, 1989:84)

1. Reading Test

The reading test is the ability to read text, process it and understand its meaning. The level understanding comes from the interaction between the word that are write and how they trigger knowledge outside the text or message.

1. Speaking Test

The speaking test is a face to face conversation with a certiflied examiner. It is as close to a real life situation as a test can get. The examiner will ask you aout familiar topics such as home. This should help you feel comfortable when speaking. Try and relax so that you can speak as naturally as possible. Take time before the test to practice speaking with a partner, friend, or teacher.

1. Listening Test

Test listening separately from speaking, since the two skills are typically exercised together in oral interaction. There are occasions, such as listening to radio or listening to song.

**Common Test Techniques**

1. Multiple Choice Items

Multiple choice can be the testing of receptive skills without requiring the test taker to produce written or spoken language (Hughes, 1989:76).

1. Short Answer Items

Items in which the test taker has to provide a short answer are common particulary in listening and reading tests (Hughes, 1989:79).

1. Gap Filling Items

Gap filling items for listening or reading work best if the missing words are to be found in text or are straightforward, high frequency words which should not present spelling problems and gap filling items also used for test of grammar and vocabulary (Hughes, 1989:80).

**Characteristics of a Good Test**

Beyond simple logic, there are technical criteria that assessment professionals use to evaluate the quality of tests and other measurement procedures. Test users often speak of the psychometric soundness of tests, two key aspects of which are reliability and validity (Cohen&Swerdlik, 2010:107)

1. Validity

Validity is the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment. According to Gronlund (in Brown, 2003:22) Validity is competency measurement tool it has function as a measuring tool. So that the measuring instrument is able to measure what should be measured. Three types of validation are important in your role as a classroom teacher: content validity, face validity, and construct validity.

1. Reliability

A reliable test is consistent and dependable. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. (Brown, 2003:20)

**Concept Related to Validity and Reliability**

**Validity**

Validity is the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment. According to Gronlund (in Brown, 2003:22) Validity is competency measurement tool it has function as a measuring tool. So that the measuring instrument is able to measure what should be measured. In general there are three types of validity, namely content validity, criterion validity, and construct validity.

1. Content Validity

Hughes states that a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. The test would have content validity only if it included a proper sample of the relevant structures. The importance of content validity: First, the great a test’s content validity, the more likely it is to be an accurate measur of what it is supposed to meansure. Secondly, such a test is likely to have a harmful backwash effect. (1989:26).

1. Criterion-related Validity

Criterion validity is a measure of how one variable or set of variables predicts on outcome based on information from other variables, and will be achieved if a set measure from a personality test relate to a behavioral. (Hughes, 1989:27).

1. Construct Validity

Test content validity and criterion-related validity provide evidence for its overall, or construct validity. However, they are not only source of evidence. Content validition of the test might confirm that thee sub-skill were well represented in the test. (Hughes, 1989:30).

1. Face Validity

A test was said to have face validity if it looks as if it measures what it is supposed to measure. Face validity is not scientific notion and is not seen as providing evidence for construct validity, yet it can be very important. (Hughes, 1989:33).

**Reliability**

A reliable test is consistent and dependable. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. (Brown, 2003:20). There are three basic strategies to estimate the reliability of most tests: the test-retest, equivalent forms, and internal consistency srategies. (Brown, 2005:193)

1. The Test-retest Reliability

Test-retest reliability is the one most appropriate for estimating the stability the test over time. The test is involved two times with the same question for the group of students, but the testing session should be far enough apart so that students are not likely to remember the items on the test.

1. Equivalent-forms Reliability

Equivalent-forms Reliability or usually called parallel-forms reliability is similar to test-retest reliabily. However, instead of administering the same test twice, the tester administers two different but equivalent tests (for example. Forms A and B).

1. Internal-consistency Strategies

Internal-consistency strategies have the advantage of estimating the reliability of a test with only one form and only one administration of that form. The easiest internal-concistency strategy to understand conceptually is called the split-half method. Split-half method is if The test is usually split on the basis of odd and even numbered items. The odd numbered and even numbere items are scored separately as though they were two different forms. There are two formulas on split-half method they are spearman-brown prophecy formula and cronbach alpha.

Among the many other variations of internal-consistency reliability, the most commonly reported are the Kuder and Richardson formula 20 (K-R20) and formula 21 (K-R21). To calculate K-R20 and K-R21, a tester only needs to know the number of items, the mean and the standard deviation of a test. To know the reliability of the teacher made test, the writer using K-R20 formula.

K-R20 = (1 - )

∑IV

St2

k

k – 1

Figure 2.1 The formula of Kuder and Richardson (K-R20)

Where: K-R20 = Kuder-Richardson Formula

k = number of item

IV = item variance

St2 = variance of the whole test

|  |
| --- |
| Criteria of reliability |
| 0.00-0.20 Not reliable  0.20-0.40 Less Reliable  0.40-0.60 Reliable enough  0.60-0.80 Reliable   * + 1. Very Reliable |

Figure 2.2 The criteria of reliability

**Methods**

This study was conducted using a qualitative and quantitative analysis method. The qualitative analysis was used to evaluate the content validity of the test. The quantitative analysis was used to evaluate the reliability of the teacher made test. The design used in this study is a qualitative descriptive.

**Sampling Technique**

The Population of this research is teacher made test at Eight grade of MTs Bustanul Ulum Kasiyan Timur - Puger to the students tested. This study will be presented a table of data that will explain the results of the classification and the percentage of those items in the MTs Bustanul Ulum Kasiyan Timur - Puger. The goal is to describe the question criteria accordance with the level of validity and reliability.

**Data Collecting Methods**

Data collecting method is a method used to obtain the things that are accurate, relevant, and appropriate to the subject matter. These data were analyzed to obtain a result that is used to draw a conclusion. Data collecting methods used in this research is to use the method of documentation.

The data to be obtained by the method of documentation in this study include:

1. Documents of syllabus;
2. File English test prepared by the teacher;
3. Recapitulation of student answer sheet’s test results;

**Item Analysis in Language Testing**

1. Norm – Referenced Item Analysis

Norm referenced item analysis involved the use of two statics: item facility and item discrimination. I will explain each of these in some detail, but I should stress at the outset that these stactical analyses are only useful insorfar as they help teacher to understandstand and improve the effectiveness of their item formats and content. Teachers should be careful to keep these statistical teqnique in perspective, remembering that the statistical are only tools for improving actual test items and are not an end in themselves. (Brown, 2005:66).

1. Item Facility Analysis

An analysis used to examine the precentage of students who correctly answer a give item. To calculate *IF,* and up the number of students who correctly answered a particular item and divide that sum by the total number of students who look the test. As an equation it would look like this.

IF = N correct

N Total

1. Item Discrimination Analysis

Item discrimination (ID) is a statistic that indicates the degree to which an item separetes the student who performed well from those who did poorly on the test as a whole. These two groups are sometimes referred to as the “high” and “low” scorers or “upper” and “lower” are proficiency students. The reosen for identifying these two groups is that *ID* allows teachers to contrast the performances of the upper students on the test with the performance of the lower students. (Brown, 2005:68-69)

ID = IF upper – IF lower

**Research Findings and Discussion**

**Validity**

The researcher using four parts of validity, they are content validity, face validity, construct validity, and criterion related validity. Where to know about content validity researcher using standard competence and based competence on the syllabus which appropriate with the question. On general result from the third question of the daily test question, the content validity is not have high validity. Face validity by looking at how the shape and appearance matter wheter the writing is near and has a sequence, and on the third question of the daily test, the layout is good. Contruct validity by selecting or observe the most difficult problem that has been done by the students and the students are not able to answer, and from the research found that most of students can’t be answer the question on the daily test 1, 2 and 3. The latter is the criterion related validity that the writer use the data items or the value obtained by the students for the measure or in value if meets the KKM standards specified by the school. And the results of the students answer is very low, that is not appropriate with the KKM on the 75 points, so the validity is not have high validity.

**Reliability**

A reliable test is consistent and dependable. If you give the same test to the same student or matched students on two different occasions, the test should yield similar results. (Brown, 2003:20). Means that from repeated calculating from the first question and the second question given the similar result or not too different. In this research used K-R20 formula. In this method the way to analyse with the devide the test arranged by the total of question, that is with analyzed each item of question.

In this research, there are 3 daily test question to analyse.daily test 1, daily test 2 and daily test 3.

1. Item analysis on daily test 1

Based on the appendix 2, researcher found the test variance (2.22), sum of the Item Variance (1.766) and number of items (18) can be substituted into the K-R20 formula. After that we can found the results on appendix 2 that have 0.67 coefficient reliability and we can said that the question of daily test 1 is Reliable, because have Reliable criteria.

1. Item analysis on daily test 2

Based on the appendix 2, researcher found the test variance (0.92), sum of the Item Variance (0.64) and number of items (18) can be substituted into the K-R20 formula. After that we can found the results on appendix 2 that have 0.222 coefficient reliability and we can said that the question of daily test 2 is less reliable, because have Less Reliable criteria.

1. Item analysis on daily test 3

Based on the appendix 2, researcher found the test variance (0.52), sum of the Item Variance (1.3928) and number of items (18) can be substituted into the K-R20 formula. After that we can found the results on appendix 2 that have -4.841 coefficient reliability and we can said that the question of daily test 3 is not reliable, because the criteria is very low.

Based on the result of the validity and reliability, the writer can see have high the teacher made test. From the result above, the test is not have high validity and scored of test not have high reliable because the scored of students on daily test 2 and daily test 3 in the less of reliability.

Furthermore, if the teacher will make an exam, the teacher must look at how to make valid and reliability examinations with the knowing criteria of valid and reliability in accordance with the standard of competency curriculum used by the school so that easier to make the teacher exam.

**Result of Interview**

The interview was held on September 23, 2015 which involved the teacher Mrs. Istiqomah, S.Pd.I. According to the result of interview it was known that teacher didn’t know about the criterion to determine a good evaluation instrument such as validity and reliability, especially on content validity because of the teacher limited to learn some technique to arrange a good evaluation instrument and standardized, and when the teacher start to make a question she didn’t matched with the content standard of curriculum on syllabus that is because when the teacher teaching in the class, she never been make the lesson plan first. And sometimes if the teacher got bad mood, she cans not make the question well. That is so disappoint for researcher, properly before the teacher teach on the class, she has to makes the lesson plan first adjust with the curriculum in used at the school, the lesson plan is very important for teacher to make the class efficient and in a fact in MTs Bustanul Ulum, the teacher didn’t does it, and about the students at MTs Bustanul Ulum, why the most of the students can not get the good mark? This phenomenon can be happen because the students can’t be understood spread all over the material which had explained by the teacher and from the question, any some tricking answer for students. That is also influenced by the students ability on English subject, more of the students at MTs Bustanul Ulum didn’t like with English subject, so they can’t be studying English happily.

**Conclusions**

Based on the data analysis which is done, so it can be concluded that on generally the item of the teacher made test on daily test 1, 2 and 3 has not fulfill the criteria of content validity. From face validity on daily tests 1, 2 and 3 the teacher using layout well. On construst validity, the teacher using different skill to make the question between daily test 1, 2 and 3. And from Criterion-related validity on daily tests 1, 2 and 3 has not fulfill standard of KKM. And for reliability has not reliable because on daily test 2 and 3 have low reliability.

In conclusion, after all careful analysis in this research, from the validity, the validity not have high validity and the items not have high reliable in the scored from the students for eight grade students.

**References**

Brown, Douglas. 2000. *Teaching by Principles an Interactive Approach to Language Pedagogy: 2nd Edition*. San Fransisco: Longman

. 2003. *Language Assessment Principle and Classroom Practice*. San Fransisco: Longman.

Cohen, R.J. & M, Swerdlik. 2010. *Psychological Testing and Assessment: An Introduction to Test and Measurement: 7th Edition.* USA: McGraw-Hill

Hughes Artur. 1989. *Testing For Language Teacher*, Cambridge: cambridge university press

Harmer, Jeremy. 2001. *The Practice of English Language Teaching: 3rd Edition.* San Fransisco: Longman

Dean Brown, James. 2005. *Testing in Language Programs: A Comprehensive Guide to English Language Assessment International Edition*: San Fransisco: Longman

Soure from Internet

<http://bsnpindonesia.org/id/wpcontent/uploads/kompetensi/PanduanUmumKTSP.pdf>

<August 15,2015>

<https://akhmadsudrajat.files.wordpress.com/2009/12/buku-saku-ktsp.pdf>

<August 15,2015>