Journal of English Education and Teaching (JEET)
e-ISSN: 2622-5867
p-ISSN: 2685-743x
Volume 3 number 4, December 2019
Page 522-532

An Analysis of Validity and Reliability of A Teacher-Made Test (Case Study at XI Grade of SMA N 6 Bengkulu)

Agung Setiabudi
English Education Study Program, Department of Language and Art
Universitas Bengkulu
budiagung663@gmail.com
Mulyadi
English Education Study Program, Department of Language and Art
Universitas Bengkulu
ladunimulyadi@gmail.com
Hilda Puspita
English Education Study Program, Department of Language and Art
Universitas Bengkulu
puspitahilda@gmail.com

Abstract

This research aimed to find out the validity and reliability of a teacher-made test in SMAN 6 Bengkulu. This research was descriptive quantitative research. The subject of this research was an English teacher-made test for eleventh-grade students of SMAN 6 Bengkulu. There were 40 items of the test consist of 35 multiple choice items and 5 essay items. There were two research instruments used in this research, observation checklist and documentation. Based on the data analysis, it was found that the percentage of the validity score was 60% or 0.60. It means that the test was valid but in the intermediate category. The r-obtained for reliability test was 0.62, and it was considered to be reliable but was in the intermediate level of reliability. From the research finding, it can be concluded that the test was valid and reliable but both were in the intermediate category. It means that the test still needs some revision and improvement to be a good, valid, and reliable test.

Keywords: *Validity, Reliability, Teacher-made test*

Introduction

In this globalization era, learning a foreign language, especially English, is becoming an important need for people to gain more competitive advantage. That is why, in Indonesia, English is the first foreign language learned to start at the elementary school level up to the university level. In conducting an effective Teaching-Learning Process (TLP), some matters should be paid attention to. They are the teacher, curriculum, syllabus, method, facility, test, etc. The test is one of the matters that will be focused on this study. Hughes (2003) stated that a test is a tool to measure language proficiency of students. Brown (2004) stated that a test is a method of measuring a person's ability knowledge, or performance in a given domain. Therefore, A test supposed to be able to measure learning outcome which distinguish the every single student's ability between students already mastered and not yet the learning material.

Tests play a vital role in the education system as they are used as tools in measurement and evaluation processes. As Mpofu (2011) states, for a teacher to be able to do his/her work effectively, he/she needs to assess the progress of his or her students from time to time. Good knowledge of where the students are and how he/she is progressing helps the teacher to effectively cater to the needs of students (Chakanyuka, 2000).

Therefore, because it is very important to measure the students' scores, a test should be valid and reliable so that the students' scores will be also valid and reliable. Based on Arikunto (2005), the concept of validity is a test that is given by the teacher should be valid. Valid test means the test can measure what it is wanted to measure, not else. For example, if the teacher wants to measure the speaking ability of the students, so the teacher should give questions orally, and the students should answer orally too. Arikunto added that "*A test is valid if it measures what it purposes to measure*". So, in other words, the test is valid if it measures what it is intended to measure.

The test which is going to be analyzed is designed by the teacher of SMA N 6 Bengkulu. The researcher will see the validity and the reliability of the test, of four eleventh grade classes that the teacher taught. The school had some facilities such as sports facilities, a natural sciences laboratory, a library, a Language Laboratory, etc. The school has used K-13 (Kurikulum 2013). It is the same as the other schools that also have used this kind of curriculum. In K-13, the government allows the teachers or the members of the committee of each school to arrange and improve the curriculum or syllabus by themselves under the coordination from the regency. If it is related to the test, the test is to measure the success of the material in the curriculum and syllabus that has been reached.

Furthermore, in formal education, a curriculum is the set of courses and their content offered at a school or university. While a syllabus is an outline and summary of topics to be covered in a course, it is often either set out by an exam board or prepared by the teacher who teaches the course and is usually given to each student during the first class session. Therefore, as in SMA N 6 Bengkulu, the tests for the final exam are made by the teacher of each subject. So, in this research, the researcher will check the English test created by the teacher to see whether it already consists of the material that is in the syllabus and curriculum of the subject.

The reason the researcher wanted to research this school is that after doing the internship, the researcher found that almost every test, most of the students had a low score. There were only a few students who had a high score. Therefore, the researcher wanted to see why the problem is happening by analyzing the validity and the reliability of the semester test in SMA N 6 Bengkulu. In this research, the researcher will see the validity and reliability of a teacher who teaches eleventh-grade students of SMAN 6 Bengkulu. The researcher will examine the validity and the reliability of the test of four eleventh grade classes that the teacher taught. Based on the problems that already explained above, the main research question addressed by this research is 'Is the teacher-made-test in Eleventh Grade of SMA N 6 Bengkulu valid and reliable?'.

Research Methodology

In conducting this research, the researcher used a descriptive quantitative research type. Quantitative research involves the collection of data so that information can be quantified and subjected to statistical treatment to support or refute "alternate knowledge claims" (Creswell, 2003). Creswell (2002) stated the researcher uses mathematical models as the methodology of data analysis. Three historical trends about quantitative research include research design, test and measurement procedures, and statistical analysis. Quantitative research also involves data collection that is typically numeric, and the researcher tends to use mathematical models as the methodology of data analysis. Additionally, the researcher uses the inquiry methods to ensure alignment with statistical data collection methodology.

. The researcher selected a set of the final test, made by the English teacher in SMAN 6 Kota Bengkulu as the subject of the research. The set of the final test chosen by the researcher in this research was a set of final test for the eleventh-grade students in SMAN 6 Kota Bengkulu in the academic year 2018/2019. This final test consists of 40 items, 35 items are multiple-choice form, and 5 items are essay question form. As for this research, the researcher only took the multiple-choice items to be analyzed because the answers can be more easily corrected than the essay form.

In conducting this research, the researcher uses the instrument to gather the data. The instruments of this research are observation checklist, documentation, and expert judgment. The data analysis of this study consist of validity and reliability. The analysis of the content validity of each test will be done by comparing the indicator in the curriculum to the content of each item of the test. Then calculate the percentage of the learning material in the content of each item. In this research, to see the consistency of the test reliability, the researcher doing the test-retest reliability. The students' worksheets from the two tests are analyzed and then correlated to obtain some description of the test items related to their reliability. In this research, Kuder Richardson's formula (KR-20) was to measure the reliability of the test and retest. The scores for KR-20 range from 0 to 1, where 0 is no reliability, and 1 is perfect reliability. If the score is the to 1, the test would be more reliable.

Findings and Discussion

Findings

In this section, the researcher presents the result of the research that has been conducted. After analyzed the data, the researcher found that from 35 items of the English final test, there were 21 items included in the indicators in the syllabus and 14 items that were not included in the syllabus. It means that 60% of the items were included in the indicators of the syllabus and 40% of the items were not included in the syllabus. The percentage can be seen in the table below.

Table 3. The Percentage of Test items in Indicators of the Syllabus.

KD	INDICATORS	NUMBER OF	TOTAL	ENTAGE
	R	ITEMS	NUMBER	
D.1	I.1 I.2 I.3 I.4	12 30 14,34 13	5	$\frac{5}{35}$% = 14.3%
D.2	I.1 I.2 I.3 I.4	9, 21, 23 27, 32 31 -	6	)% = 17.1 %
D.3	I.1 I.2 I.3 I.4	- - 7 8	2	)% = 5.7 %
D.4	I.1 I.2 I.3	15 16, 17 18, 19	5	)% = 14.3 %
D.5	I.1 I.2	26, 28 33	3	)% = 8.6 %
TAL			**21**	**)% = 60 %**

From the table above, it can be seen that 21 items were included in the indicators from five KD in the syllabus. In each KD, several indicators need to be achieved by the students at the end of the study.

Then, the table below was the percentage of the item test that was not valid. It was the test items that were not in the indicator of the syllabus.

Table 4. The Percentage of Test Items That were not Included in Indicators of the Syllabus

KD	DICATO R	NUMBER OF ITEMS	TOTAL NUMBER	CENTAGE
-	-	?,3,4,5,,6,10,11, ,22,24,25,29,35	14	$\frac{14}{35}$)% = 40%
TOTAL			**14**	$\frac{14}{35}$)% = 40%

From the table above, as we can see that there were 40% of the items that were not included in the indicators of the syllabus.

In the reliability, the researcher conducted the test twice using the same test to the same students to see the consistency of the test reliability. The first was conducted in December 2018, and the second was conducted in May 2019. The reliability of each test was measured by using the Kuder Richardson's formula (KR 20).

Based on an interpretation of the reliability coefficient adapted from Allen and Davis (1978), the final English test made by the teacher in SMA N 6 Bengkulu for Eleventh-grade class that conducted in December 2018 was moderately reliable as it was found that the score obtained by the first test was 0.60. Then, it was found that the score obtained by the second English final test made by the teacher in SMA N 6 Bengkulu for Eleventh-grade class that conducted in May 2019 was 0.665. Therefore, it was in the moderate reliability category too.

In brief, the results of the reliability of the teacher-made test of SMA N 6 Bengkulu both in December 2018 and in May 2019 were moderately reliable. The researcher then calculated the score from time 1 and time 2 using Pearson's correlation and found the stability of the test stability. The analysis then found the result of the reliability coefficient was 0.62. Based on the criteria of reliability coefficient adapted from Allen and Davis (1978), the result was in the moderate reliability category.

Discussion

Test Validity

A test is valid if it measures what is supposed to measure. So, in other words, the test is valid if it really measures what it is intended to measure. In this research, the researcher intended to find out the validity of the test. Afterwards, it can be concluded which items that can still be used, can be used with revision, or should be dropped.

After analyzing the data, the researcher found the result of the test validity. Briefly, the result showed that the total percentage of the items that was in the indicators were 60% or 0.60. It means that the test fulfill the indicators in the syllabus that should be achieved by the students in the end of the study. Based on the validity category adapted from Putri (2009), the result was in the moderate validity category. From those result, it can be concluded that the English final test made by the teacher in SMA N 6 Bengkulu was valid but in the moderate level. It means that the test was valid. It was in the average level of validity. The test was considered valid and still can be used, but it was recommended to be revised. It was supported by Henning (1987) that said validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is intended to measure.

Hence, the 40% or 0.40 invalid items need to be eliminated or revised and the activity should be truly conducted by the teacher in order to be suitable with normal validity index of a high-quality test.

If we look at the previous studies that already mentioned in chapter II, most of the previous studies had the same result concerning the teacher-made test validity. One of them was by Sugianto (2008) with the research title 'Analysis of Validity and

Reliability of English Formative Tests'. The results of the reasearch showed that the teacher-made test that being investigated in the study was valid too. So, it can be concluded that eventhough a test was made by the teacher themselves there was a possibiity that the test was valid. It can happens if the teacher made it based on the material in the curriculum learned by the students. A report by Newell (2002) asserts that teacher-made tests usually measure only a limited part of a subject area, they do not cover a broad range of abilities and they rely too heavily on memorized facts and procedures.

Test Reliability

Reliability is the quality of being consistent and trustworthy. A test will be *reliable* when it gives the same repeated result under the same conditions moreover, Worthen (1993) state that reliability refers to the consistency of the scores obtained. This research aim to find the validity and the reliability of a teacher-made test in SMAN 6 Bengkulu, therefore, beside analyzing the validity of the test the researcher also analyzing the reliability of the test.

In analyzing the test reliability, the researcher used the KR-20 and Pearson Correlation formula. The result from KR-20 showed that the score of the test reliability was 0.60 for the first test and 0.665 for the second test. Based on the reliability categorization by Allen and Davis (1978), Anthony (1983) and Worthen (1993), the score was in the moderate reliability category. Moreover, from the Pearson correlation formula, the score obtain was 0.62 and also considered to be moderately reliable as it was also belongs to moderate reliability category.

From the result above, it can be conluded that the reliability of the test was questionable or doubtful because all the score obtained were in the moderate reliability category. If we look at the previous studies on chapter II, 2 of 3 studies found that the teacher-made test was also moderately reliable. It can happened because of some factors in which different school may have different factor affecting the reliability of the test.

There were three variables that can affect the validity and reliability of teacher-made tests: the test taker, the environment and the test. The student responses to these questions should be consistent. A student who gets one of these questions right and the other wrong is not a reliable test taker and should not be used to assess the validity of the test itself (Cassel, 2003). The testing environment is another variable associated with the validity of teacher-made tests. If the testing environment is distracting or noisy or the test-taker is unhealthy, he or she will have a difficult time remaining consistent throughout the testing process (Griswold, 1990).

Even though actions ought to be taken to ensure that the testing environment is comfortable, adequately lit with limited interruptions (Griswold, 1990), these factor and the former one are largely aspects of test administrative procedures that are external to the test itself. This is because even in contexts where the characteristics of the test taker and the environment are well taken care of, it emerges that individual difference in performance will still be recorded. This means that the third intrinsic variable affecting reliability and validity of teacher-made tests no matter the characteristics of the test-taker and the environment is the quality of tests themselves. The length of tests, use of Bloom's taxonomy in test item construction and prior training of teachers on test construction to enable the teachers to design items that address various cognitive levels of thinking as per the Bloom's taxonomy across the curriculum will all affect the validity and reliability of a given test.

Conclusion

From the result of this research, it can be concluded that the questions of the test were related to the indicators of the syllabus used by the teacher. The result showed that the teacher-made final test of SMA N 6 Bengkulu was valid, but the validity was in the moderate validity level, with a total percentage of 60%. Then, from the calculation, it was found that the coefficient of reliability of the test items is 0.62. This result of reliability of the test showed that the teacher-made final test of SMA N 6 Bengkulu was reliable, but based on the categorization of reliability, the reliability was in the moderate reliability level category.

In general, the researcher concluded that the items in this final English test for the eleventh-grade students of SMA N 6 Bengkulu were moderately valid and reliable.
Based on the explanation and the conclusion above, the researcher intended to give some suggestions. The teacher should pay attention more to the test items. It should be appropriately matched to the instructional objective/standard competence stated in the curriculum or the indicator stated in the syllabus used in the school. Furthermore, when the teacher found an item's test was not in the indicator of the syllabus, the teacher should revise the items that can still be used in the test.

Moreover, the researcher hopes that the result of this teacher-made test analysis could be used as an example in analyzing other test items, and encourages another researcher to do further research on a similar object. Further, further researchers were expected to conduct a research with similar topic which should be done with greater population in order to gain a wider generalization.

References

Anthony J. (1983). *Educational Tests and Measurement an Introduction*. New York: Harcourt Brace Jovanovichi, inc.

Arikunto, S. (2005). *Dasar-dasar evaluasi pendidik*an. Jakarta: Bumi Aksara.

Brown, H. D. (2004). *Language Assessment, Principles and Classroom Practices*. San Fransisco: Longman.

Cassel, R. (2003). *Confluence is a primary measure of test validity and it includes the creditability of test taker*. College Student Journal, 37:348-353.

Creswell, J. (2003*)*. *Research design: Qualitative, quantitative and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Creswell, J. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Merrill Prentice Hall.

Chakanyuka, S. (2000). *Measurement and evaluation*. Harare: Zimbabwe Open University.

Griswold, P. A. (1990*) Assessing relevance and reliability to improve the quality of teacher-made tests*. NASSP Bulletin, 76:18-24.

Heaton, J.B. 1991. *Writing English Language Tests*. New York: Longman inc.

Henning, G. 1987. *A Guide to Language Testing*. Mass: New Bury House Publishers.

Hughes ,A. (2003). *Testing for Language Teachers. Retrieved from https://www.ukessays.com/essays/education/definition-of-test-types-of-test-educationessay.php on Januari 17 2019.*

Mpofu, B. (2011). *Formative evaluation versus summative evaluation*. Harare: Longman

Slameto. (2001). *Evaluasi pendidikan*. Jakarta: Bumi Aksara.

Sugianto, A. (2008). *Analysis of Validity and Reliability of English Formative Test.*

Worthen, B. R., Borg, W. R., and White, K. R. (1993). *Measurement and evaluation in the schools*. White Plains, NY: Longman.