*Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*
||Volume||3||Issue||2||Pages|| 22 – 38 ||2017||
Available online at:http://ejournal.kopertais4.or.id/mataraman/index.php/efi |P-ISSN: 2460-0938; E-ISSN: 2460-2604|

# VALIDITY AND RELIABILITY OF ENGLISH SUMMATIVE TEST FOR SENIOR HIGH SCHOOL

**Aris Sugianto**
*aris.sugianto@iain-palangkaraya.ac.id*

Institut Agama Islam Negeri Palangka Raya, Indonesia
*Jl. G. Obos Komplek* Islamic Centre *Palangka Raya, Indonesia*

**Article History:**
Received:
December 5, 2016

Revised:
July 12, 2017

Accepted:
December 22, 2017

**Corresponding Author:**
*aris.sugianto@iain-palangkaraya.ac.id*

**Abstract:** This study aims to analyze statistically the validity and reliability of English summative test for the second semester of the tenth graders of SMAN 2 Palangka Raya in academic year 2015/2016. The writer used descriptive quantitative approach to find the result. In analyzing the validity, the writer used Point-biserial correlation formula, while to analyze the reliability, the writer used K-R20 formula. The validity was analyzed based on test item, and the validity of the whole test was determined based on the percentage of all valid items. The result was that the English summative test was valid and reliable. Specifically, from 50 items of questions, 32 items (64%) were valid and 18 items (36%) were invalid. Interpreted from the 64% of valid items, so overall the summative test was valid in substantial level. The summative test was also reliable. The coefficient of reliability was .907. Therefore, the reliability was in the level of excellent reliability.

**Keywords – Analysis, Validity, Reliability, English Summative Test, Characteristics of Good Test**

## INTRODUCTION

Test is one of good important components in teaching and learning process. There are the components which relate each other actually in teaching and learning process. They are goal, activity, and evaluation. Goal is the purpose of the lesson, activity is the process or activity in classroom itself, and evaluation is the procedure to measure the success of the goal and the activity. Test is a kind of evaluation which is as instrument of measurement of teaching and learning process, instrument to see the ability/ achievement of students, or instrument to take educational decision.

It is known that a test should fulfil standard or characteristics of a good test. Based on theories, there are five characteristics of a good test. They are: 1) the test should be valid, 2) the test should be reliable, 3) the test should be objective, 4) the test should be practicable, and 5) the test should be economic(Djiwandono, 2008; Sudijono, 2011). From the five characteristics of a good test, the two firsts are the most important; they are the test should be valid and reliable.

Validity is the extent to which the test measures what it is wanted to measure. In other words, a valid test can really measure what it is supposed to measure. For instance, if the test is supposed to measure speaking ability, so the test is constructed, conducted, answered orally. Rajhy(2014)stated that the term validity refers to the extent to which the test measures what it says it measures. In line with Ary, Hughes (2003, p. 26) also stated "A test is said to be valid if it measures accurately what it is intended to measure". Consequently, validity refers to the suitability between a test as an instrument of measurement and the domain of what it is supposed to measure.

The extent of validity can be analysed logically or empirically. The kind of validity analysed logically is called as logical validity, while the validity analysed empirically is called as empirical validity (Sudijono, 2011). Logical validity is resulted from the process of thinking logically. So a test if it has been supposed that has fulfilled logical validity, it means that the test rationally has been able to measure what it should measure. Empirical validity is resulted from the correlation between the test and the empirical data. Empirical data are obtained from experiences which are interpreted into numbers or scores.

Logical validity can be divided into two types; they are content validity and construct validity. Content validity is the extent to which the test measures the materials that has been taught and programmed in syllabus. So, in content validity, items should be representatives of whole materials. Hughes (2003, p. 26) stated "A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned". Rajhy (2014) stated that content validity is the extent to which the selection of tasks one observes in a test taking situation is representative of the larger set of tasks of which the test is assumed to be a sample, so in the other words that a test should be a representative sample of the teaching/instructional contents as defined and covered in the curriculum.

Actually, content validity is automatically fulfilled when the construction of the test finished, because logically the teacher should construct the test based on the materials in syllabus. So it can be concluded that the condition of logical validity does not need to test/analyze, but automatically gained after the construction finished. But if it is needed, to

test/analyse the content validity, generally the researcher/analyzer only needs to compare each item of test to the materials programmed in syllabus. Analysis can be done before or after the test conducted.

Construct validity is the extent to which the test measures aspects of thinking (Sudijono, 2011). Aspects of thinking refers to psychological theory such mentioned by Benjamin S. Bloom; they are cognitive domain, affective domain, and psychomotoric domain. Cognitive, affective, and psychomotoric aspects are constructed as have been established in Specific Instructional Objective of the learning. This is in line with Rajhy (2014) who stated that construct validity is the relation between a test and the Psychological abilities it measures.Construct validity should also automatically fulfilled when the construction of the test finished, because logically the teacher should construct the test based on aspects of thinking which have been established in Specific Instructional Objective of the learning.

Besides validity analysis by using analysis logically and empirically above, there is a technique that also can be used; it is item analysis. This technique is called as test item validity analysis. The technique of this analysis is by correlating the score of each item to the total score. The validity of whole test can be determined by the value of the percentage of all valid items.

Another important characteristic of a good test is that the test should be reliable. Reliability refers to the consistency of scores. A test would be considered as a good test and credible as a measurement of learning achievement if the result/score of test is consistent in how many times the test conducted. Consistence does not mean the scores have to be the same, but it can be stable changing of scores. A test is considered reliable if the result (scores) is approximately the same repeatedly (Sugianto, 2016a). Brown (2005, p. 175)stated "Test reliability is defined as the extent to which the result can be considered consistent or stable". Also Rajhy (2014) stated that the results of test should be consistent where they remain stable and the test should not produce different results when it is used in different days. So, a test that is reliable will yield similar results with similar group of students took the same test on two occasions, and their results are roughly the same. For instance, if a test is give twice, in first test student A gets score 60 and student B gets score 80, so the scores will be indicated consistent if in the second test, the result/scores of students A gets such 70 and student B gets 90 or close to those scores. The scores of the second test can be conformably increased or decreased. If the result/scores have been consistent, it means the test is reliable, and it is credible to be instrument of measurement of learning achievement.

In measuring the extent of reliability of test, many techniques and formulas can be used. Techniques and formulas are also related to the types of data. So, techniques, formulas, and data are closely related each other.

In case of this study, summative test is a type of test analysed in terms of validity and reliability. Summative test is kind test which is conducted after all units of learning material finished to be learned or taught. In Summative test (assessment), the teacher wants to find out what the students can remember about the course material so that a mark can be determined (Qu & Zhang, 2013). Based on the explanation above, the result of summative test can be used to get scores and to determine the educational decision.

Summative test is one of tests which are categorized based on the function in teaching and learning process. Actually there are some kinds/terms of tests categorized based on the function, such as pre-test, post-test, formative test, and including summative test (Djiwandono, 2008). Pre-test is type of test conducted before the teaching and learning process. So, the purpose of pre-test is to know the prior ability of the students before treatment. This prior ability is the knowledge obtained from the previous grade. Schalich(2015) stated that the pre-test is assessing the student's knowledge of the previous grade and then progresses based on what the students should learn by the time of the next assessment period.

Post-test is type of test conducted after the teaching and learning process. So, the purpose of post test is to know the achievement of the students after treatment. In a research, pre-test and post-test are usually used to measure the effect of any treatment. So pre-test and post-test are used as a method of research. The deviation between the result of pre-test and post-test is analysed to see the level of significance by using certain statistical formula (Djiwandono, 2008).

Formative test is type of test conducted after every unit finished. It is in order to see achievement of students in every unit learned. Formative test is types of assessment which is as a part of instructional process where the result can provide information needed to adjust teaching and learning while they are happening, so the adjustments helps to ensure students achieve the learning goals within a set time frame (Garrison & Ehringhaus, 2007).

While, summative test is type of test conducted after all unit finished. It is in order to see achievement of students for all unit learned. In Indonesia, summative test is conducted in the end of every semester. The same with the result of formative test, the result of summative test also can provide the information related the elements of teaching and learning process including curriculum, materials, teaching method, exercise and tests have been used. So, summative test is a part of comprehensive evaluation of teaching

program. As part of comprehensive evaluation of teaching program, so the materials for constructing the summative testinvolve all materials those have been taught from the beginning of semester to the end of the semester(Djiwandono, 2008).

In this study, the writer focuses on the English summative test for the second semester of the tenth graders of SMAN 2 Palangka Raya in academic year 2015/2016. This study is aimed to give information and as reference for the teachers and researchers. For the teachers, the result of this research can be valuable information about the condition of summative test made by the English teacher, especially in terms of statistical validity and reliability. So, in the future the teacher can construct the test well and better. While for the researcher, this study can be reference to enrich knowledge about theory and practice of related topic; test and its analysis. The writer chose SMAN 2 Palangka Raya because this school is the model of best school in Palangka Raya. SMAN 2 Palangka Raya is a school with accreditation A. the writer assumed if the result of analysis of teacher-made test is good, the school can be the model for the other schools, but if the result is not good, it will be a big question in other schools.

There are some previous studies related to this study. The first is a research written by Sugianto (2016) entitles"An Analysis of English National Final Examination for Junior High School in Terms of Validity and Reliability". The study was to analyse the validity and reliability of the English National Final Examination for Junior High School. The study was analysed by using the descriptive method. Content validity was analysed logically (logical validity) by comparing the materials in syllabus to the items of the test, and construct validity was analysed by comparing the indicators in syllabus to the items of the test. The reliability was analysed by using Kuder Richardson Formula (KR-20). The result of the study showed that the English National Final Examination for Junior High School was valid and reliable. The content validity showed 100% valid, and the construct validity showed 100% valid. While the reliability showed coefficient 0.89, and it meant reliable. So, the English National Final Examination for Junior High School has fulfilled the characteristics of a good test. The study investigated about validity and reliability of a test where the validity was analysed logically (type of logical validity) which used qualitative method. It is rather different with this recent study where the validity is analyzed statistically (type of empirical validity).

The second previous study is written by Setiyana (2016) entitles "Analysis of summative tests for English". The study aimed to analyse the quality of summative tests for English at MAN Boarding School Meulaboh I in terms of validity, reliability, difficulty index, discrimination index, and the effectiveness of distractors. Content analysis was

employed in this study. Two techniques were carried out to collect the data, namely a checklist and document analysis. The data from the checklist was analysed using statistical procedures and the data from the document analysis was analysed using *Anates* software version 4. The results showed that the validity of the English summative tests at MAN Meulaboh I was on average either sufficient or poor since the percentages were below 72%.Secondly, the tests had a high and consistent degree of reliability. The index of difficulty was above 70%. Thirdly, 60% of the difficulty index in the test of the first grade, 48% in the second grade, and 8% in the third grade test were accepted. Fourthly, more than half of the discrimination index was good. In detail, good in the discrimination index of the test was 76% in the first grade, 56% in the second grade and 72% in the third grade. Finally, the effectiveness of distracters in the English summative test in the first grade was 53%, in the second grade was 67% and in the third grade was 50%. In the study, the analysis was complete to see the quality of a test. The writer analysed the English summative test based on validity (logical and empirical validity), reliability, difficulty index, discrimination index, and the effectiveness of distracters, while this recent study focuses on the analysis of the validity statistically (empirically) and reliability. Nevertheless, the result can be reference of the next study.

The third previous study is written by Haryudin(2015) entitles "Validity and reliability of English summative tests at junior high school in West Bandung". The study was purposed to measure the validity and reliability of English summative test items for the third grade of Junior High School in West Bandung. Thestudywas categorized as quantitativedescriptive analysis.The results, there were 21 items (70%) of the test regarded valid because the value of correlation coefficient result (r) was greater (>) than table value (r-table) = 0.213 for the 5% level. Meanwhile, the number of correlation coefficient (r) by using KR-20 of the test was in the amount of 0.71. The correlation number of 0.71 lied between the interval 0.70-0.90 with a high interpretation. It can be concluded that the English Summative test has good validity and high reliability. Based on significance and method of the study, this recent study is the same. The different is the place and the objective of research. If the study was conducted in Junior High School of West Bandung, this recent study is conducted in Senior High School of Palangka Raya.

The fourth previous study is written by Agustito(2012) entitles "An Analysis of English National Final Examination (UN) for Junior High Schools in Kurun Viewed from School-Based Curriculum (KTSP)". The writer finds that there is a previous studies dealing with analysis of English National Final Examination to criteria of a good test. The analysis is to match whether the English National Final Examination is matching with the

competencies and materials in English syllabus of Junior High Schools in Kurun, and the result is it is matching.

The fifth previous study is also written by Sugianto (2011) entitles "Analysis of Validity and Reliability of English Formative Tests". The study was conducted in order to analyze the validity and reliability English formative tests made by the English teacher of grade VIII of SMPN-4 Mentaya Hulu. The validity covered the content and construct validity. The writer applied descriptive method in conducting the study. The population and sample of this study was the English teacher-made tests (formative tests) for the grade VIII of SMPN-4 Mentaya Hulu in the first semester of academic year 2008/2009. The writer analysed two English formative tests as the representative of all English formative tests in the first semester of academic year 2008/2009. The result of the data analysis, the writer found that the English formative test conducted on October 2008 belongs to high validity in its content validity and excellent validity in its construct validity. The English formative test conducted on November 2008 belongs to low validity in its content validity and high validity in its construct validity. While the English formative test both conducted on October 2008 and November 2008 are unreliable. The same with the first previous study, the study investigated about validity and reliability of a test where the validity was analysed logically (type of logical validity) which used qualitative method. It is rather different with this recent study where the validity is analysed statistically (type of empirical validity).

The sixth previous study is written by Fauzi (2011) entitles "An analysis of the content validity of the English summative test for the second grade of Madrasah Tsanawiyah Salafiyah Bedahan Kota Depok". The research aimed to find the empirical evidence of the English summative test validity, in this case especially content validity. The summative test was made by the professional team (KKM) for the second grade students of Madrasah Tsanawiyah Salafiyah Bedahan. The summative test consisted of 45 items where 40 items were multiple choice questions and 5 essay questions. The research used descriptive method. The results of the research showed that items of the English summative test for even semester of the second grade students in Madrasaha Tsanawiyah Salafiyah Bedahan have bad content validity. It meant the materials of the English summative test were not appropriate to the recommended English syllabus. The study was analyzing especially on the content validity the English summative test for the second grade of Madrasah Tsanawiyah Salafiyah Bedahan Kota Depok. So, it was only comparing the suitability of the content of English summative items to the material in syllabus of the school. The analysis was conducted logically (types of logical validity) and qualitatively.

The seventh previous study is written by entitles "An Analysis of Validity of English Summative Test Constructed by the Teachers for the Seventh Grade Students of SLTPN-1 Pahandut". The results of Marleni's study showed that the content validity belongs to the poor qualification, the construct validity belongs to the very low qualification, the criterion-related validity belongs to the good validity level, and the wash back validity belongs to the low qualification. The study discussed especially on the validity and the result was such not good. It is questionable if the test was used to be summative test while it is know that the result of summative test usually used to determine educational decision. It can harm the students at the school.

The eight previous study is written by Claritha (2006)entitles "An Analysis on the Summative Test made by the Teacher of SMP Katolik Palangkaraya". While some results of Claritha's study showed that the content validity belongs to the high qualification, the construct validity belongs to the fair qualification, the criterion-related validity belongs to the very low category, the reliability of the Multiple Choice Question (MCQ) belongs to the reliable level, and the reliability of the Essay Question belongs to the unreliable level.

From the studies above, it can be concluded that the tests constructed by the English teachers (teacher-made tests) still have many problems. The analyses of quality of test still become informative needs for English teachers and the schools to pay more attention to improve their knowledge and ability in constricting a good test based the characteristics of a good test such mentioned by Djiwandono (2008) that besides validity and reliability, there are some other aspects that should be analysed to prove that the test has fulfilled the standard of good quality (the characteristics of a good test) such as mentioned in the elements of test item analysis; they are item difficulty, item discrimination, and distracter analysis. So, to make the test qualified, these elements (validity, reliability, item difficulty, item discrimination, and distracter analysis) should be in good level (high level).

## METHOD

This research used descriptive quantitative method on processing the data. It described summative test as it was. The analysis was processed through obtained scores. The population of this study was the English summative test made by the English teachers of SMAN 2 Palangka Raya. The sample was the English summative test for the second semester of the tenth graders of SMAN 2 Palangka Raya in academic year 2015/2016. There were some classes that got summative test at the semester such as Class X IPS 1, Class X IPS 2, Class X IPS 3, Class X MIPA 6, Class X MIPA 7, and Class X MIPA 8.

The writer chose randomly a summative test for Class X MIPA 6 as the data source of the sample. So, the data sources were from the question sheet and answer sheets of 38 students of Class X MIPA 6. The summative test consisted of 50 items.

Validity was analysed by correlation technique and appropriate formula. It was analyzed based on each item of the summative test. So theoretically it belonged to test item validity. The appropriate formula was determined based on the types of data (Sugianto, 2016c).The correlation formula used Point-biserial. The writer used Point-biserial correlation formula because the technique was by correlating dichotomous data to interval data. Theoretically, if the correlated data are between the scores of each item to the total score of the test where the score of each item consists of dichotomous data (objective test scores) and the total score is interval data (total score of the objective test), the appropriate formula that should be used to get accurate result is Point-biserial Correlation formula (Sugianto, 2016b). Sudijono (2011, p. 185) stated that if the variable I is discreet data or dichotomous data, and the variable II is continuous data, the appropriate correlation technique to be used is Point-biserial Correlation. Brown (2001) also stated the point-biserial correlation coefficient (symbolized as $r_{pbi}$)is a statistical measurement used to estimate the degree of relationship between a naturally occurring dichotomous nominal scale (individual item)and an interval (or ratio) scale. Point-biserial is one of statistical tools (see the Table 1)

**Table 1. Types of Correlation Coefficients**

| **Correlation Coefficient** | **Types of Scales** |
|---|---|
| Pearson product-moment | Both scales interval (or ratio) |
| Spearman rank-order | Both scales ordinal |
| Phi | Both scales are naturally dichotomous (nominal) |
| Tetrachric | Both scales are artificially dichotomous (nominal) |
| Point-biserial | One scale naturally dichotomous (nominal), one scale interval (or ratio) |
| Biserial | One scale artificially dichotomous (nominal), one scale interval (or ratio) |
| Gamma | One scale nominal, one scale ordinal |

These are the formulas of Point-biserial Correlation:

Formula I

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

Formula II

$$r_{pbi} = \frac{M_p - M_t}{S_t}\sqrt{\frac{p}{q}}$$

Where:
$r_{pbi}$ : Point-biserial correlation coefficient
$M_p$ : mean on the whole test for those students/testee who answered correctly (coded as 1s)
$M_q$ : mean on the whole test for those students/testee who answered incorrectly (coded as 0s)
$M_t$ : mean of total scores
$S_t$ : standard deviation for whole test
*p* : proportion of students who answered correctly on the whole test
*q* : proportion of students who answered incorrectly on the whole test

*p* and *q* can be calculated by the following formula:

$$p = \frac{Np}{N}$$

$$q = 1 - p$$

Where:
N*p* : Number of students who answered correctly on the whole test
N : Number of whole students

The correlation coefficient (r-observed/$r_{11}$) was interpreted by consulting with r-table. The coefficient correlation was consulted to the r-table based on the value of Degree of Freedom (df) with significant level 5%. For the correlation, the formula of Degree of Freedom isdf=N-2. If r-observed ≥ r-table, it means the item is valid, and if the r-observed < r-table, it means the item is invalid. The whole test could be interpreted based on percentage of all valid items. The extent of validity could use the following general interpretation range:

.80(80%) – 1 (100%) : High to Very High
.60 (60%) – <.80 (80%) : Substantial
.40 (40%) – <.60 (60%) : Moderate

.20 (20%) – <.40 (40%) : Low

.00 (0%) – <.20 (20%) : Negligible

Reliability was analysed by using Kuder-Richardson formula; it is K-R20. The formula of K-R20 is often used to analyse the reliability of test if the data is dichotomous data such for objective test score (usually 1 for correct, 0 for incorrect) and the technique is single test – single trial approach. Below is the K-R20 formula:

$$r_{11} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{S_t^2}\right)$$

Where:

$r_{11}$ : reliability coefficient of test
k : number of items
1 : constant number
$S_t^2$ : total variance
$p$ : proportion of students who answered correctly on the whole test
$q$ : proportion of students who answered incorrectly on the whole test

Generally, the reliability coefficient can be interpreted by the criteria:

$r_{11} \geq .70$ : reliable

$r_{11} < .70$ : unreliable

General Interpretation by the level:

.90 and up : excellent

.80 - .89 : good

.70 - .79 : adequate

below .70 : may have limited applicability

## FINDINGS AND DISCUSSIONS

### Validity

In this study, the validity of the summative test was analysed in each item. The analysis was by correlating the score of item to the total score. The score of item was in form of dichotomous data and the total score was in form of interval data, so the writer used Point-biserial correlation to analyse the data. The coefficient was interpreted by r-table with Degree of Freedom (df)=N-2 and Significant level on 5%. The whole test was interpreted based on the sum of valid items.

Here is the result:

**Table 2. Distribution of Item Validity**

| | | |
|---|---|---|
| Numbers of Items | 50 | |
| Valid Items | 32 | 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 20, 21, 22, 23, 26, 29, 30, 31, 32, 33, 35, 38, 39, 41, 45, 46, 48, 49, 50 |
| Invalid Items | 18 | 1, 2, 12, 17, 18, 19, 24, 25, 27, 28, 34, 36, 37, 40, 42, 43, 44, 47 |

Where:

| | |
|---|---|
| N | 38 |
| df=N-2 | 36 |
| r-table 5% | 0.320 |

From the Table 1., it can be seen, there 32 items are valid and 18 items are invalid. The interpretation was consulted to the value of r-table at significant level 5%; it was 0.320.So, the valid items were determined if$r_{11} \geq 0.320$, while the invalid items were determined if$r_{11} < 0.320$.

The statistical analysis found some fundamental cause of invalid items. Besides the coefficients were below 0.320, there some items were undefined. These undefined coefficients were caused if the items could be answered correctly by all the students. It meant the items were too easy. Good items should not be too easy and not too difficult. In the theory of item difficulty, based on Djiwandono (2008, p. 219), on the extreme level where the items are able to be answered by all students correctly or the items are not able to be answered by all students, it means the items are not effective. The items where no one can answer correctly or the items can not discriminate the ability among the students (all students can answer correctly), means the items are not useful. So, the items are categorized as invalid. These types of invalid items (undefined coefficient) happened to 25, 27, 34, 36, and 37. The items could be answered by all students correctly or too easy.

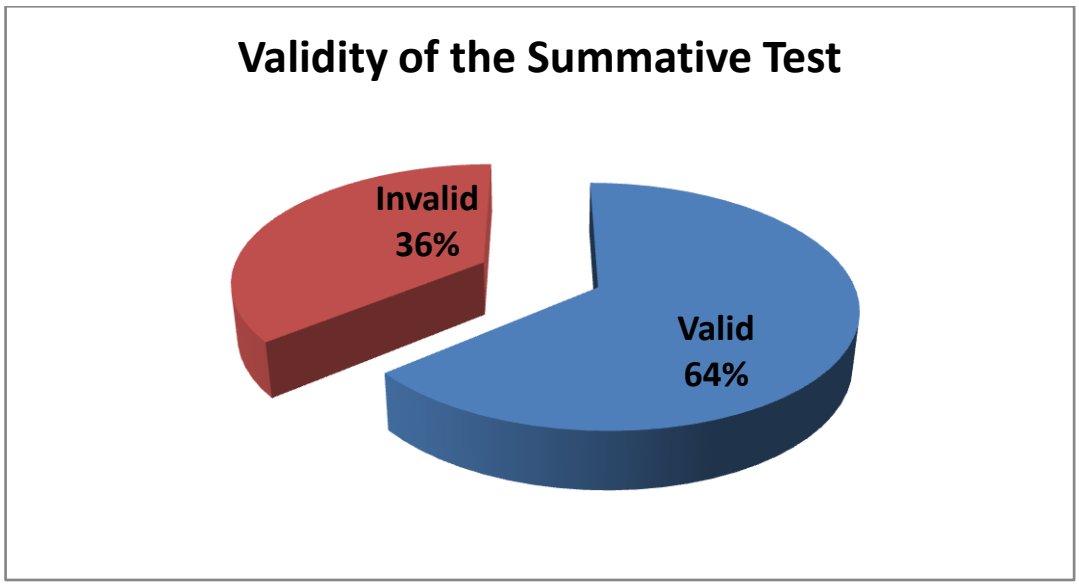


**Figure 1. Chart of Validity of Summative Test**

From the Figure 1., it shows that 64% items were valid and 36% were invalid. From the percentage of valid items, the extent of validity of whole test could be determined. The validity of whole test was interpreted based on the following general interpretation range:

.80(80%) – 1 (100%) : High to Very High
.60 (60%) – <.80 (80%) : Substantial
.40 (40%) – <.60 (60%) : Moderate
.20 (20%) – <.40 (40%) : Low
.00 (0%) – <.20 (20%) : Negligible

The whole test was categorized as valid test if the value of percentage was started from .60 (60%) and more than it. So, since the percentage was 64%, it could be categorized that the Summative test was valid in the level of substantial validity.

**Reliability**

Reliability refers to the consistency of scores if the test is given to the test-takers two occasions or more. Consistent scores do not mean the scores have to be the same exactly, but it can be approximately the same or stable changing of scores. The reliability level is important to be analysed in order to see whether the test is credible to be an instrument of assessment. The credibility is indicated by the consistency of the scores. In this study, since the form of summative test was multiple choice questions, the reliability was analysed by using K-R20 Formula. The calculation was as follow:

k : 50 items

$\sum pq$ : 5.52

$S_t^2$ : 49.64

$$\boldsymbol{r_{11} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{S_t^2}\right)}$$

$$\boldsymbol{r_{11} = \frac{50}{50-1}\left(1 - \frac{5.52}{49.64}\right)}$$

$$\boldsymbol{r_{11} = 1.02(0.889)}$$

$$\boldsymbol{r_{11} = 0.907}$$

So, the reliability coefficient was 0.907. The coefficient was interpreted by criteria:

$r_{11} \geq 0.70$ : reliable

$r_{11} < 0.70$ : unreliable

General Interpretation by the level:
.90 and up : excellent
.80 - .89 : good
.70 - .79 : adequate
below .70 : may have limited applicability

So that, based on the criteria, since the coefficient was 0.907, it was interpreted that the summative test was reliable in the level of excellent reliability.

## CONCLUSIONS AND SUGGESTION

A good test should fulfil the characteristics of a good test. It should be valid and reliable. Based on this study, it can be concluded that statistically the summative test for the second semester of the tenth graders of SMAN 2 Palangka Raya in academic year 2015/2016 is valid and reliable. It means that the summative test fulfil the characteristics of a good test. Thus, the summative test can be the instrument of measurement of teaching and learning process, instrument to see ability/ achievement of students, or instrument to take educational decision.

In this study, the validity was analysed based on items. So every item was analysed one by one to see the extent of validity. The coefficient of item validity was interpreted by using r-table on criterion 0.320. If the coefficient is 0.320 or more, the item is valid, less than 0.320 is invalid. It was found that 32 items were valid and 18 items were invalid. The

invalid items were also influenced by the difficulty level of items. There were 5 items were too easy so all students can answer correctly. The results/ coefficients were undefined, so they were also categorized as invalid.

Based on the percentage of valid items, the validity of whole test could be known. The validity of whole test was considered by the value of percentage of all valid items. From 50 items of test, 32 items were valid, and the rests were invalid. It meant that 64% items were valid. So, overall, summative test was interpreted as valid test.

Reliability refers to the consistency of scores. If the test is conducted repeatedly and the scores are consistent, it means the test is credible to be used as instrument of educational measurement. In this study, the reliability was analysed by using K-R20 formula. The result of analysis by the formula was gotten coefficient 0.907. Based on the coefficient, it was interpreted that the summative test was reliable in the level of excellent reliability.

Besides validity and reliability, actually there are some other aspects that should be analysed to prove that the summative test has fulfilled the characteristics of a good test (good quality test) such as mentioned in the elements of item analysis which consists of item difficulty, item discrimination, and distracter analysis (Djiwandono, 2008).

Furthermore, in the term of validity, not only statistical analysis, but also it can be analysed logically such as mentioned in the types of validity. But in this occasion the writer only has opportunity to analyse the statistical validity and the reliability. It will be worth consideration and suggestion if the next researcher or next author's writing will analyse these other aspects.

From 50 items of questions, it is only 64% of valid items. Even thought 64% is in level of substantial, but for the level of school with accreditation A, according to the writer it is not expected result. The school is expected to be role model for other schools which have not got the accreditation. To be the role model, it should be good result or better than this, at least it is on high level (≥80%). From this view, so it is suggested to the teachers of the school to improve their knowledge about how to construct a good test. To improve the knowledge, the school can delegate the teachers to follow some workshops or by holding some workshops and obligating the teachers to participate.

## REFERENCES

Agustito, A. (2012). *An Analysis of English National Final Examination (UN) for Junior High Schools in Kurun Viewed from School-Based Curriculum (KTSP)* (Unpublished). Palangka Raya University.

Brown, J. D. (2001). Point-biserial Correlation Coefficients. *JALT Testing & Evaluation SIG Newsletter*, *5*(3), 12–15.

Brown, J. D. (2005). *Testing in Language Programs*. New York: McGraw-Hill.

Claritha, F. (2006). *An Analysis of the Summative Test Made by the Teacher of SMP Katolik Palangka Raya* (Unpublished). Palangka Raya University.

Djiwandono, S. (2008). *Tes Bahasa*. Jakarta: PT. Indeks.

Fauzi, S. (2011). *An Analysis of the Content Validity of the English Summative Test for the Second Grade of Madrasah Tsanawiyah Salafiyah Bedahan Kota Depok*. UIN Syarif Hidayatullah Jakarta, Jakarta. Retrieved from http://repository.uinjkt.ac.id/dspace/bitstream/123456789/3026/1/SALMAN%20FAUZI-FITK.pdf

Garrison, C., & Ehringhaus, M. (2007). *Formative and Summative Assessments in the Classroom*. Retrieved from http://ccti.colfinder.org/sites/default/files/formative_and_summative_assessment_in_the_classroom.pdf

Haryudin, A. (2015). Validity and Reliability of English Summative Tests at Junior High School in West Bandung. *P2M STKIP Siliwangi*, *2*(1), 77–90.

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge: Cambridge University Press.

Marleni, M. (2006). *An Analysis of Validity of English Summative Test Constructed by the Teachers for the Seventh Grade Students of SLTPN-1 Pahandut* (Unpublished). Palangka Raya University.

Qu, W., & Zhang, C. (2013). The Analysis of Summative Assessment and Formative Assessment and their Roles in College English Assessment System. *Journal of Language Teaching and Research*, *4*(2). https://doi.org/10.4304/jltr.4.2.335-339

Rajhy, H. A. A. (2014). Five Characteristics of a Good Language Test. *National Journal of Extensive Education and Interdisciplinary*, *2*(4), 61–66.

Schalich, M. E. (2015). *Analysis of Pre-Test and Post-Test Performance of Students in a Learning Center Model at the Elementary School Level*. Dominican University of California, California. Retrieved from http://scholar.dominican.edu/cgi/viewcontent.cgi?article=1181&context=masters-theses

Setiyana, R. (2016). Analysis of Summative Tests for English. *English Education Journal*, *7*(4), 433–447.

Sudijono, A. (2011). *Pengantar Evaluasi Pendidikan*. Jakarta: Rajawali Pers.

Sugianto, A. (2011). Analysis of Validity and Reliability of English Formative Tests. *Journal on English as a Foreign Language*,*1*(2), 87–94.

Sugianto, A. (2016a). An Analysis of English National Final Examination for Junior High School in Terms of Validity and Reliability. *Journal on English as a Foreign Language*, *6*(1), 31–42.

Sugianto, A. (2016b). Technique and Procedure of Statistical Analysis of Test Item Validity of Objective English Language Test. In *21st Century English Language Teaching* (Vol. 2, pp. 5–18). IAIN Palangka Raya. Retrieved from https://www.researchgate.net/profile/Karya_Ilmiah_Tadris_Bahasa_Inggris/publication/313045322_Proceeding_21st_Century_English_Language_Teaching/links/588eaee6a6fdcc8e63cac784/Proceeding-21st-Century-English-Language-Teaching.pdf#page=6

Sugianto, A. (2016c). Types of Variable Data (Discreet Variable and Continuous Variable). In *Role of International Languages toward Global Education System* (pp. 5–7). IAIN Palangka Raya. Retrieved from https://www.researchgate.net/profile/Karya_Ilmiah_Tadris_Bahasa_Inggris/publication/311570913_Proceedings_of_International_Conference_Role_of_International_Languages_toward_Global_Education_System/links/584d871308ae4bc899330a31.pdf#page=10

**Contributor's Biodata**

Aris Sugianto, M.Pd. is a full time lecturer of State Islamic Institute of Palangka Raya (IAIN Palangka Raya), Central Kalimantan, Indonesia. At 2014, he was a part time lecturer of this institution. He has been as full time lecturer (state lecturer) from 2015 until present. He earned his Undergraduate degree (2009) and Master degree (2013) in ELT from State University of Palangka Raya. His expertise is English Language Evaluation. Email: aris.sugianto@iain-palangkaraya.ac.id, HP: 082154409206