VALIDITY AND RELIABILITY OF ENGLISH SUMMATIVE TESTS AT JUNIOR HIGH SCHOOL IN WEST BANDUNG

Acep Haryudin

Program Studi Pendidikan Bahasa Inggris, STKIP Siliwangi

ABSTRACT

This study is purposed to measure the validity and reliability of English summative test items for the third grade of Junior High School in West Bandung. This research is categorized as quantitativedescriptive analysis because it is intended to describe the difficulty level, discriminating power, distracters effectiveness, validity and reliability of the English Summative Test. The finding of this study are that there are 16 items (53,33%) regarded as easy test items in difficulty level that range from 70-1.00 and 12 items (40%) of total items have satisfactory discriminating power range from 0,20-0,40. In the term of the effectiveness of distractor, 17 items (56,7%) of the distractors are poor. Therefore, this tes has easy difficulty level, satisfactory discriminating power and poor distractors. Moreover, there are 21 items (70%) of the test regarded valid because the value of correlation coefficient result is greater (>) than table value (rt) = 0.213 for the 5% level. Meanwhile, the number of correlation coefficient (r) of the test is in the amount of 0.71. The correlation number of 0.71 lies 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.

Keywords: Validity, Reliability, English Summative Tests

ABSTRAK

Penelitian ini bertujuan untuk mengukur Validitas dan Reliabilitas item tes sumatif bahasa Inggris yang diujikan kepada siswa SMPN kelas tiga di Bandung Barat. Penelitian ini dikategorikan sebagai analisis deskriptif karena ini bertujuan untuk menggambarkan tingkat kesulitan, daya pembeda, efektifitas pengecoh, validitas dan reliabilitas item tes bahasa Inggris. Temuan didalam penelitian ini adalah bahwa terdapat 16 item (53,33%) dianggap sebagai tes item yang mudah didalam tingkat kesulitan dari 70-1.00 dan 12 item (40%) dari total item memiliki daya pembeda yang memuaskan dari 0,20-0,40. Kaitannya dengan epektifitas pengecoh, 17 item (56,7%) adalah pengecoh lemah. Dengan demikian, tes ini memiliki tingkat kesulitan yang mudah, daya pembeda yang memuaskan, dan pengecoh yang lemah. Selain itu, terdapat 21 item (70%) dari tes dianggap valid karena nilai dari hasil korelasi koefisiensi adalah lebih besar (>) dari nilai tabel (rt)=0,213 untuk tingkat 5%. Sementara, jumlah korelasi koefesiensi (r) dari tes adalah sejumlah 0,71. Korelasi nomor 0,71 berada diantara interval 0,70-0,90 dengan penafsiran/interpretasi tinggi. Hal ini dapat disimpulkan bahwa Tes Sumatif Bahasa Inggris memiliki validitas yang baik dan realibilitas yang tinggi.

Kata Kunci: Validitas, Reliabilitas dan Tes Sumatif Bahasa Inggris

A. INTRODUCTION

The advancement of nations is measured by the strength of their first-rate educational system, where qualified persons are prepared to be at a high degree of efficiency and creativity in order to be able to develop the community. Besides, they are prepared to have high flexibility to develop themselves and to keep up with changes and developments of the period. This depends on the quality of the means of evaluation and measurement which help in making objective decisions on a scientific basis.

Evaluation is a comprehensive process of collecting and constructing data to make judgments about a particular programme or group of people. It includes collecting, analyzing and interpreting information about teaching and learning in order to make conversant decisions that enhance student achievement and the success of educational process (Allen, 1998: 170).

Measurement is the first step in the process of evaluating student's achievement. So it has its own tools that are used to evaluate the students' performance in form of marks, then the measurement starts by comparing the marks with particular standard previously set by the curriculum development centre. After processes of interpretation, the marks are used to identify the strengths and weaknesses. The final process comes (the treatment process) where the strengths are strengthened and the weaknesses are treated whether for the students, teachers, content or the teaching methods (Kopriva, 2008: 55).

Measurement, evaluation and achievement tests are different concepts, whether in definition or purpose, yet they are interrelated interdependent processes and they complement each other. This relation can be described as a system that we cannot do evaluation without them since measurement precedes evaluation. Episcopal (2008) believes that achievement test is one of the evaluation tools by which the teacher attains scores that represent measurement. Ultimately, the evaluation is based on the interpretation of these scores.

The first thing to be considered while preparing the test is the aim of the test, where each test has a specific aim. For example, the English test emerges from the objectives of the curriculum being taught. After getting done with the aims of the test, a new stage starts which is the stage of applying the table of specification.

Furthermore, going back to the teachers' methods being applied in designing English tests at junior high school, it is clear that the majority of them are not aware of the test standard rules. The test standards are mainly represented in comprehensiveness of the given material in the test. In other words, the taught material is necessary for the test validity and reliability. This is considered a real problem especially when it comes to English tests at public junior high school for the following reasons:

1. Not all of the content standards are available in the English tests at junior high school.
2. Not all of the general standards of designing test available in the English tests at junior high school.

This study comes to focus on the different sides of the problem to study the quality standards, to measure the extent of its validity and reliability in the English summative test.

Based on the statement above, the writer is interested in analysing the test items made by the English teacher at public junior high schools in west Bandung and the esearcher did the research under the tittle "Validity And Reliability Of English Summative Tests At Junior High Schools In West Bandung".

Based on the background, the writer identifies some problem are:

a. Is there a team of language experts to create English test items?
b. How valid are the English test items at junior high school?
c. How reliable are the English test items at junior high school?
d. How does teaching-learning process of English take place?
e. How does English teacher measure test items?
f. Are the English summative tests at junior high school valid?
g. Are the English summative tests at junior high school reliable?
h. Does the test measure what it is intended to measure?

To make this study easier to understand, the writer limits the study as follows:

1) The research focused only on the validity and reliability of English Summative Test for the third year students at oddsemester 2013/2014.
2) The test which is analyzed is English Summative Test for the thirdyear students at odd semester, 2013/2014 academic year.
3) The research focused only on the third year students ofpublic junior high schools in west Bandung.

The following research questions are generated from the main question:

a) Do the English Summative tests at public junior high schools in west Bandung have good validity?
b) Do the English Summative tests at public junior high schools in west Bandung have good reliability?

The purposes of this study are:

(1) To identify the validity of the English Summative tests at public junior high schools in west Bandung.

(2) To identify the reliability of the English Summative tests at public junior high schools in west Bandung.

It's expected that the study will provide some useful inputs for both the pupils and the English teachers, especially the English teacher at Junior High School. And the writer expects that this study could provide the first steps and ways for further study.

B. REVIEW OF LITERATURE AND METHODOLOGY

1. Review of Literature

a. Test

Allan (1998:183) defines test as an instrument of evaluation by which we attempt to measure learner performance. Besides it has a physical existence and operates within specific time frames, seeking an accurate prediction about the basis of relatively small samples of performance in the case of such an enormously complex thing as language.

Brown (2004: 3) defines tests as methods of measuring a person's ability, knowledge, or performance in a given domain. Most common forms of tests include fill–in–the blanks, sentence completion, open answers, and multiple choices.

b. Significance of Test

Although most students and some teachers abhor tests, the need for testing is great. There are many reasons for giving a test, of course it's important to get a vivid idea about the real function of the test which is give to students. The value of test stems from the role it plays to evaluate the students' behavior, motivation, and to encourage and promote them.

1) Achievement

Achievement is the action of accomplishing something. A test may be used to evaluate a student' achievement of what should have been taught but not inevitably what has actually been taught. Every student likes to know what he/she has achieved, to what extent he/she has achieved and where he/ she stands amongst his classmate (Allison, 1999: 97).

2) Motivation

Motivation is the psychological feature that arouses a person act toward a desired goal.On the other hand a high score makes him feel pleased as it is said "success leads to further success" (Conner, 1999: 127).

3) Encouraging students

Regrettably it is true that many students study only for tests. In fact the great majority of students do not study unless a test is declared (Conner, 1999). Therefore, tests are probably the only support for them to work hard. In other words, tests encourage student to take their learning seriously.

4) Diagnosis

Sometimes it is necessary to diagnose (conclusion following the test) problems and difficulties in managing a function, a concept and perceptions involving language skills and sub skills. In other words, some tests are designed to discover students' weaknesses. So a remedial work could be prepared to deal with such weaknesses.

5) Self-evaluation

Corner (1999: 166) defines self-evaluation as an evaluation of oneself or one's performance in relation to an objective standard. Tests are now and again needed for the teacher to evaluate his own teaching methods. The feedback he/ she get from the tests assists him/ her very much to amend the way he teaches.

6) Experimentation

Tests may be used successfully in educational experiments in order to determine a certain technique of teaching or a certain hypothesis. In this regard a pre-test and a post-test are usually given to an experimental and control clusters.

7) Promotion and Advancement

Some tests are sometimes designed to decide which students are to be promoted from a grade to a upper one. Without testing, promotion will be involuntary or impressionistic. (Harlen, 2007).

8) Parents Information

Tests can give parents information about their children's levels (e.g. how they are moving ahead, the areas of weaknesses and distinction and the type of help they require). (Harlen, 2007).

c. Test Criteria

A good test should possess the following qualities:

1) Validity

Validity is an expansive construct that engages making appropriate interpretations and uses of scores or test information (Morgan & Anderson, 2008). One aspect of validity is the extent to which the content of a test is representative of the curriculum or concept that is being measured.

a) Content Validity

One of the most significant points in test validation is exploring whether the test is related to a given area of content or ability. In the matter of language tests, one of the principals that concern content validity is the extent to which a test measures a representative sample of the language in question (Robertson and Nunn, 2008: 162).

b) Criterion Validity

Exploring the validity of a test by means of external criteria is seen as essential by many scholars. Criterion-related proof exhibits a relationship between test scores and some criterion which is believed to be also an indicator of the ability tested.

c) Concurrent Validity

This refers to how well scores on a new test mach the scores procured by other previously validated measures of equivalent skills (Alshumaimeri, 1999: 5).

d) Construct Validity

The major concern of language test designers is whether test performance truly reflects language abilities or not. Construct validation helps to validate the extent to which a testee's performance on a particular test can be indicative of his/her fundamental competence. Construct validity (Alshumaimeri, 1999: 5), refers to "the extent to which performance on tests is compatible predictions that we make on the basis of a theory of abilities, or constructs".

e) Predictive Validity

This refers to the relationship between scores achieved by a measure such as a proficiency test and the language performance of the students while they use the language in the actual world (Alshumaimeri, 1999: 5).

f) Face Validity

According to (Waine. and Braun, 1988: 23) the term face validity is related to the tests look, reasonability and quality. It concerns the people perception of the test in general.

2) Reliability

Brown (2004) states that, "a reliable test is consistent and dependable". This means that an elementary concern about the development and the use of language tests is its reliability, that is, the constancy of the test as a measure. Reliability refers to the consistency of the examination scores. Also, it refers to the scope to which the test produces consistent results if different markers mark it.

a) Types of Reliability

(1) Test-retest reliability

The same test is administered twice and a correlation calculated between the scores on each administration (Fulcher and Davidson, 2007: 105). The scores from Time 1 and Time 2 can then be correlated in order to evaluate the test for stability over time.*Example:* A test designed to assess student learning in psychology could be given to a group of students twice, with the second administration perhaps coming a week after the first. The obtained correlation coefficient would indicate the stability of the scores.

(2) Parallelforms reliability

Two forms of the same test are produced, such that they test the same construct and have similar means and variances. The correlation between the scores on the two forms is taken as a measure of reliability (Fulcher and Davidson, 2007: 105). The scores from the two versions can then be correlated

in order to evaluate the consistency of results across alternate versions.*Example:* If you wanted to evaluate the reliability of a critical thinking assessment, you might create a large set of items that all pertain to critical thinking and then randomly split the questions up into two sets, which would represent the parallel forms.

(3) Inter-rater reliability

Inter-later reliability occurs when two or more scorers yield inconsistent scores of the same test, possibly lack of attention to scoring criteria, inexperience, inattention, or even preconceived biases (Brown, 2004: 21). Inter-rater reliability is useful because human observers will not necessarily interpret answers the same way; raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed.*Example:* Inter-rater reliability might be employed when different judges are evaluating the degree to which art portfolios meet certain standards. Inter-rater reliability is especially useful when judgments can be considered relatively subjective. Thus, the use of this type of reliability would probably be more likely when evaluating artwork as opposed to math problems.

(4) Internal consistency reliability

Internal consistency reliability is a measure of reliability used to evaluate thedegree to which different test items that probe the same construct produce similarresults (Phelan and Wren2005-06).

3) Practicality

The last quality that a good test should have is practicality or usability. In selecting test and other instrument, practical considerations can not be neglected. These are some factors relevant to the practicality when selecting tests (Allison, 85).

(a) Ease of administration

The administrability of evaluation devices refers to the ease and accuracy with which the directions to pupils and evaluator can be followed (Remmers, 1960: 126)

(b) Time reqired for administration

The test lengt is direcly related to the reliability of a test, so the availability of enough time should be taken. “a safe procedure is to allot as much time as is necessary to obtain valid and reliable results” (Tinambunan, 1988).

(c) Ease of interpretation and application

If the test is interpereted correctly and applied effectively, teacher can make accurate educational decisions about students performance.

(d) Availability of equivalent or comparable forms

Equivalent test measure the same aspect and is alike in content, level of difficulty and other characteristics. It is useful if teacher wants to remove the factor of memory when retesting students on the same domain. Comporable forms are especially useful in measuring the progress of basic skills.

(e) Cost of testing

The factor of the cost is actually not really important in selecting test. Testing is relatively inexpensive. However, the point is the test should be as economical as possible in cost.

d. Test Types

Traditionally, five types of test are recognized [although the precise labels may diverge from one tester to another.

1) Progress Tests

This type of tests measures how well learners have absorbed the material [or the skills] taught in class and how well they have developed in a specified area. As a result the center of attention is given to the short-term objectives [such as Unit Two: the Past Tense or Unit Four: Expressions of Future Time (Hahn 2007: 30). These tests are regularly written by the teacher to answer questions like: (1) Have the students acquainted the target language well? (2)Have I taught it efficiently? (3) Can we move to the next part of the course? Progress tests are frequently given to stimulate learners and to strengthen learning.

2) Achievement Test

At the end of a course, and possibly at one or two other points during the course, the learners are examined in what they have learned from the course. This may have the use of examining the

effectiveness of the course as much as testing the learners (Mullane: 2002).

3) Short-term Achievement Assessment

At regular periods during the course, the learners may be observed to see what they are learning from the course. These parts of observing may take the type of weekly tests, the keeping of achievement records such as graphs of reading speed, charts of writing improvement and self-assessment records. This short-term assessment can take place on a daily or weekly basis. It is called "achievement" assessment because it examines items and skills obtained from the course (Macalister, 2010: 12).

4) Proficiency Tests

Proficiency Tests are different in that they are not usually established upon a particular curriculum, but are used to measure achievement in relation to a definite [future] task that the candidate may be required to do at a following point of time. For example, the test may set out to decide whether the candidate has sufficient English to follow a special course for which the medium of instruction is English, or to do a job that requires the use of English. These tests infrequently take into account any curriculum that a student may have followed: they are concerned with future [possible] performance rather than past achievement. They are frequently administered to learners with varied language learning backgrounds. The common factor is the purpose to which the language is to be set (Mullane, 2002: 79).

5) Placement Tests

The learners are assessed at the beginning of a course to see what level of the classes they should be in. The aim of this testing is to make sure that the Tests which are constructed purposely to collect evidence about ability to learn are defined as aptitude or ability tests. Results on such tests are used to anticipate future success based upon success in the particularly selected tasks in the aptitude test. Regularly these tasks differ from the usual school learning requirements and rely to some extent on learning beyond the school curriculum (Ross, 2005: 22). Of course, teaching students the test items and the corresponding answers may bring about an increase in score without actually changing a student's (actual) aptitude. The course is not going to be too easy or too difficult for the learner (Macalister, 2010: 109).

6) Aptitude or Ability Test

Tests which are constructed purposely to collect evidence about ability to learn are defined as aptitude or ability tests. Results on such tests are used to anticipate future success based upon success in the particularly selected tasks in the aptitude test. Regularly these tasks differ from the usual school learning requirements and rely to some extent on learning beyond the school curriculum (Ross, 2005: 22). Of course, teaching students the test items and the corresponding answers may bring about an increase in score without actually changing a student's (actual) aptitude.

7) Practical Tests

In some senses an essay test is a sensible task. The essay item requires a candidate to perform. This performance is intended to express meaning in a practical sense by writing prose to an agreed format. However, the term "practical test" goes beyond performance and other tasks used in traditional pencil-and-paper examinations. The term may refer to practical tasks in trade subjects (such as woodwork, metalwork, shipbuilding, and leathercraft), in musical and theatrical performance, in skills such as swimming or gymnastics, or may refer to the skills required do laboratory or field tasks in science, agriculture, geography, environmental health or physical education (Mullane, 2002:103).

8) Diagnostic Tests

It allows teachers to identify any areas of weakness or complexity, thus they can then plan and put into practice a suitable remedial teaching program. They may be used to assess the knowledge and skills of learners in specific aspects of language before the start of a course [and accordingly may be used for placement as well as course design purposes] (Sangster and Overall, 2006:21).

2. Research Methodology

This research is categorized as descriptive analysis because it is intended to describe the difficulty level, discriminating power, distracters effectiveness, validity and reliability of the English Summative Test. The participants of this research

were the third grade students of SMP Negeri 02 Ngamprah, and SMP Negeri 03 Padalarang in West Bandung academic year 2013/2014. The total number of the students were 344 students, which were divided into eight classes. The writer took 25% of the total number of the second grade students as a sample because the population is more than 100 subjects (Arikunto, 2006: 134). The writer used a random sampling to get the students' answer sheet. The writer took the students' answer sheet rendomly so every student in population is considered as the same subject.

a. Population and samples

The population of this research is the third grade students of SMP Negeri 02 Ngamprah, and SMP Negeri 03 Padalarang in West Bandung academic year 2013/2014. The total number of the students are 344 students, which are divided into eight classes. The writer took 25% of the total number of the second grade students as a sample because the population is more than 100 subjects (Arikunto, 2006: 134). The writer used a random sampling to get the students' answer sheet. The writer took the students' answer sheet rendomly so every student in population is considered as the same subject.

b. Technique of Collecting Data

The procedures are used in doing the study as follows:

1) Asking for permission from principal and English teacher at PublicJunior High Schools in West Bandung to do the research, and asking for the English Summative test paper, the answer keys and the students' answers sheet.
2) Checking the answer keys of the test whether they are correct or not for each items, and the pupils' answer sheet.
3) Tabulating and calculating how many pupils who answer correctly from each group (upper and lower groupt and then put it in the format of tabulation of item analysis.
4) Calculating the index of difficulty and discriminating power of each item.
5) Analyzing which items need to be revised or maintained based on the definition.
6) Analyzing the distactors of the items which must be revised.
7) Analyzing the validity and reliability of the test items.

c. Technique of Data Analysis

When a test is too easy so that almost everyone gets all the items right or conversely, almost everyone gets all the items wrong, it does not have a certain quality. Quality of a test depends on the quality of the items. If the items of a test are good, it can be said that the test is also good, on the contrary it can be said that the test is bad, if the items of test are bad. In order to know the quality of test item, it should be analyzed. Ngalim Purwanto (1994:118) points out, "the aim of item analysis is to find the good test items and the bad test items and why the items are categorized as a good item test or bad item test (tujuan khusus dari item analysis ialah mencari soal test mana yang baik dan mana yang tidak baik, dan mengapa item atau soal itu dikatakan baik atau tidak baik."

In doing the item analysis, there are two points that should be examined. An item analysis has some benefits for both pupils and teachers. J. Stanley Ahman and Marvin D. Glock (1967: 190) explain "By evaluating the data from Item analysis, they can detect learning difficulties of individual pupils or the class as a whole, and consequence more suitable remedial programme. Studying the strenghts and weaknesses of pupils' achievement will also help teachers to evaluate more accurately effectiveness ovarious parts of the learning situation."

In accordance with the importance of using item analysis to guide teachers in making preparation of better tests in the future and improve student's learning that have stated above, here are the more detailed discussion of those.

C. FINDINGS AND DATA ANALYSIS

1. Findings

a. Difficulty Level

The difficulty level of an item is indicated by the number of students who answer the items correctly. The index of difficulty of an item shows how easy or difficult the particular item is. In this respect, the item should not be too easy or too difficult; there should be balance between them. If the test is too easy, the learners will not be motivated to answer the test. On the other hand, If the test is too difficult, the students will be frustated because they do not know how to answer.

After analysing the data,the researcher describes the difficulty level of each items in the form ofdata tabulationandcalculation. the result of thedifficulty level of English summativetest listedin table4below:

Table 4. Classification of Items Based on The Proportion of Difficulty Level

No	Index of Difficulty	Total Number	Percentage
1	Easy	1	3,33 %
2	Moderate	13	43,33 %
3	Difficult	16	53,33 %

b. Discriminating Power

A good item should discriminate between the smarter students from the poor ones. The discrimination powershould be able to differentiate between students from the upper group and those from the lower group. It can be concluded if the test items are answered correctly by the upper groups and answered incorrectly by the lower gloups, the test is good because it can distinguish between those.

The reason for identifying these two groups is that descriminating power allows teacher to contrast the performance of the upper group students on the test with of that lower group students. To do this, teacher or test maker can compare the number of students from the upper and lower group who answered the item correctly.

Based onthe analysis ofEnglish Summative test items the final result of descriminating power is showen in the table 5, as follows:

Table 5. Classification of Items Based on The Proportion of Descriminating Power

No	Discriminating Power	Total Number	Percentage
1	Very Poor	3	10 %
2	Poor	11	36,67 %
3	Saticfactory	12	40 %
4	Good	4	13,33 %
5	Excellent	0	0 %

c. Effectiveness of Distractor

The effectiveness of distractor is a procedure specifically related to multiple-choice item. Distractions function to divert students from the correct answer if they do not know which is correct. So, it is important to evaluate the quality of each distractor in multiple-choice test. Moreover, the primary goal of distractor efficiency analysis is to examine the degree to which the distractors are attracting students who dont know the correct answer. A good distractor will attract more students from the lower group and the upper group.

Based on the calculation of the number of testee who chose alternative answers of the five choices A, B, C, D, and E, the result of English Summative test the data presented the final result of distractor effectiveness is shown in table 6, as follows:

Table 6. Classification of Items Based on The Proportion of Distractor Efficiency

No.	Effectiveness of Distractor	Number	Percentage
1	Effective	**13**	**43,3** %
2	Ineffective	**17**	**56,7** %

d. Validity of Test

To analyse the validity of English Summativetestitems, the researcher has to refer to appendix2 constituting tableanalysisthat functions to find out: M_p, M_t, SD_t,pandq. According to Sudijono (2003: 187-189) there are the stepsto be takenin analyzing thevalidity of thetestitems, as follows:

Step I is preparing calculation table to analyzethe validity oftest items number1 to30(see appendix 2).

Step II is finding the meanof thetotalscore, namely M_t using the following formula:

$$M_t = \frac{\sum X_t}{N}$$

Known : Σ Xt = 1724 and N = 86

$$M_t = \frac{\sum X_t}{N} = \frac{1724}{86} = \mathbf{20,05}$$

Step III is finding the total ofstandarddeviation, namely SD_t, using the following formula:

$$SD_t = \sqrt{\frac{\sum X_t^2}{N} - \left(\frac{\sum X_t}{N}\right)^2}$$

Known : $\Sigma X_t^2 = 36148$, Σ Xt = 1724 and N = 86. So :

$$SD_t = \sqrt{\frac{\sum X_t^2}{N} - \left(\frac{\sum X_t}{N}\right)^2}$$

$$SD_t = \sqrt{\frac{36148}{86} - \left(\frac{1724}{86}\right)^2}$$

$$SD_t = \sqrt{420,33 - 402,0025}$$

$$SD_t = \sqrt{18,3275}$$

$$SD_t = \mathbf{4,28}$$

StepIV is findingand calculating M_p for test itemsnumber1to 30usingthe follwing formula:

Mp = Total score of testee who answered correctly / Number of testee who answered correctly

For calculating M_p, the researcher refers to table 7, as follows:

Table 7. Calculation for Finding M_p of Test Item Number 1 to 30

Item No	Total score of testee who answered correctly	Number of testee who answered correctly	*Mp*
1	1536	76	20,211
2	1173	58	20,224
3	1137	53	21,453
4	451	21	21,476
5	1351	63	21,444
6	1428	67	21,313
7	1324	62	21,355
8	610	29	21,034
9	1083	51	21,235
10	1294	62	20,871
11	1199	57	21,035
12	1545	75	20,600
13	684	32	21,375
14	1379	64	21,547
15	940	43	21,860
16	1322	62	21,323
17	944	43	21,953
18	965	44	21,932
19	1546	75	20,613
20	1322	62	21,323
21	1369	66	20,742
22	1132	55	20,582
23	1568	78	20,103
24	1355	64	21,172
25	1415	69	20,507
26	906	44	20,591
27	1061	52	20,404
28	1044	47	22,213
29	1458	70	20,829
30	1607	80	20,088

Step V is calculating correlation

Coefficient r_{pbi} of test item number 1 to 30 using the formula:

$$r_{pbis} = \frac{M_p - M_t}{SD_t}\sqrt{\frac{p}{q}}$$

(Sudijono, 2003: 185)

Wherein:

r_{pbis}= point biserial correlation coefficient is representing the strength of the correlation between variable 1 and variable 2, which in this regard is considered to be validity coefficient.

M_p= mean score that the testee has for the test item answered correctly.

M_t= mean score of the total score.

SD_t= standard deviation of the total score.

p= proportion/ the number of testee who answered the test items anlyzed correctly.

q= proportion/ the number of testee who answered the test items anlyzed incorrectly.

After performing step four, and finding the value of M_p, the writer then makes calculation to earn point- biserial correlation (r_{pbis}) which is thevalue ofthe validitythat every test item has.

In the interpretation of this provision db of (N-nr) is used, namely = 86- 2 = 84 (Sudijono, 2003: 190). The degree of freedom of 84 is then consulted with the table value of "r" moment product. So the results are as follows: at the 5% significance level (r_t) = 0.213 at the 1% significance level (r_t) = 0.278. If the value (r_{pbbi}) of correlation coefficient result is greater (>) than table value (rt) = 0.213 for the 5% level, the result obtained is significant, this means that the test items are regarded valid. If the value (*rpbis*) of the correlation coefficient is smaller (<) than table value (rt) = 0.213 for the 5% level, then the level obtained is non-significant. This means test items are invalid.

The calculation of correlation coefficient of point biserial is shown in the table 8, as follows:

Table 8. Calculation of Correlation Coefficient of Point Biserial (R_{pbis}) in Analyzing The Validity of Test Item Number 1 to 30

No	M_p	M_t	SD_t	p	Q	$\sqrt{\frac{p}{q}}$	r_{pbis}	$r_{tabel\ 5\%}$	Interpretation
1	20,211	20,05	4,28	0,884	0,116	2,761	0,104	0,213	INVALID
2	20,224	20,05	4,28	0,670	0,330	1,425	0,058	0,213	INVALID
3	21,453	20,05	4,28	0,616	0,384	1,267	0,415	0,213	VALID
4	21,476	20,05	4,28	0,244	0,756	0,568	0,189	0,213	INVALID
5	21,444	20,05	4,28	0,733	0,267	1,657	0,540	0,213	VALID
6	21,313	20,05	4,28	0,779	0,221	1,877	0,554	0,213	VALID
7	21,355	20,05	4,28	0,721	0,279	1,608	0,490	0,213	VALID
8	21,034	20,05	4,28	0,337	0,663	0,713	0,164	0,213	INVALID
9	21,235	20,05	4,28	0,593	0,407	1,207	0,334	0,213	VALID
10	20,871	20,05	4,28	0,721	0,279	1,608	0,308	0,213	VALID
11	21,035	20,05	4,28	0,663	0,337	1,403	0,323	0,213	VALID
12	20,600	20,05	4,28	0,872	0,128	2,610	0,335	0,213	VALID
13	21,375	20,05	4,28	0,372	0,628	0,770	0,238	0,213	VALID
14	21,547	20,05	4,28	0,740	0,260	1,687	0,590	0,213	VALID
15	21,860	20,05	4,28	0,500	0,500	1,000	0,423	0,213	VALID
16	21,323	20,05	4,28	0,721	0,279	1,608	0,478	0,213	VALID
17	21,953	20,05	4,28	0,500	0,500	1,000	0,445	0,213	VALID
18	21,932	20,05	4,28	0,510	0,490	1,020	0,449	0,213	VALID
19	20,613	20,05	4,28	0,872	0,128	2,610	0,343	0,213	VALID
20	21,323	20,05	4,28	0,721	0,279	1,608	0,478	0,213	VALID
21	20,742	20,05	4,28	0,767	0,233	1,814	0,293	0,213	VALID
22	20,582	20,05	4,28	0,640	0,360	1,333	0,166	0,213	INVALID
23	20,103	20,05	4,28	0,907	0,093	3,123	0,039	0,213	INVALID
24	21,172	20,05	4,28	0,744	0,256	1,705	0,447	0,213	VALID
25	20,507	20,05	4,28	0,802	0,198	2,013	0,215	0,213	VALID
26	20,591	20,05	4,28	0,510	0,490	1,020	0,129	0,213	INVALID
27	20,404	20,05	4,28	0,605	0,395	1,238	0,102	0,213	INVALID
28	22,213	20,05	4,28	0,547	0,453	1,099	0,555	0,213	VALID
29	20,829	20,05	4,28	0,814	0,186	2,092	0,381	0,213	VALID
30	20,088	20,05	4,28	0,930	0,070	3,645	0,032	0,213	INVALID

Based onthe analysis ofEnglishsummative test items for grade IX of SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang Bandung at odd semester2013/2014the result isshown in table5. Referring to the table 5 above, the validity of test items is stated in table9, as follows:

TABLE 9. Validity of Test Item

No	Category	Number of Test Item	Percentage
1	Valid	21	70 %
2	Invalid	9	30

e. Reliability of Test

Toanalyse thereliability of English Summative testfor grade IX of SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang Bandung at odd semester 2013-2014, the researcherrefer to appendix 1andthentune into the table of test reliabilityanalysis in appendix2wherethe datahas

been calculated to find out: ΣX_t, ΣX_t^2, p_i, q_i, dan $\Sigma p_i q_i$.

Furthermore, the researcher perform the analysis of test reliability with the following steps:
Step I is calculating multiple choice test items number 1 to 30 in the form of item analysis tabulation (appendix 2).
Step II is finding the total variance of (S_t^2) using the formula: [8]

$$S_t{}^2 = \frac{\sum X_t{}^2 - \frac{(\sum X_t)^2}{N}}{N}$$

Referring to appendix III, known that: $\Sigma X_t = 1724$; $\Sigma X_t^2 = 36148$; and N = 86. So:

$$S_t{}^2 = \frac{\sum X_t{}^2 - \frac{(\sum X_t)^2}{N}}{N}$$
$$= \frac{36148 - \frac{(1724)^2}{86}}{86}$$
$$= \frac{36148 - 34560{,}18605}{86}$$
$$= \frac{1587{,}81395}{86}$$
$$= \mathbf{18{,}463}$$

Step III is making calculation to determine reliability of the test using the formula K-R 20: [9]

$$r_{1\,1} = \left(\frac{n}{(n-1)}\right)\left(\frac{S_t{}^2 - \sum p_{i\,.}\, q_i}{S_t{}^2}\right)$$

Referring to the appendix 2 and the calculation above, it is known that:
n = 30; $S_t^2 = 18{,}463$; $\Sigma p_i q_i = 5{,}79692$. So:

$$r_{1\,1} = \left(\frac{n}{(n-1)}\right)\left(\frac{S_t{}^2 - \sum p_{i\,.}\, q_i}{S_t{}^2}\right)$$
$$= \left(\frac{30}{(30-1)}\right)\left(\frac{18{,}463 - 5{,}79692}{18{,}463}\right)$$
$$= \left(\frac{30}{29}\right)\left(\frac{12{,}66608}{18{,}463}\right)$$
$$= (1{,}034)\ (0{,}686)$$
$$\boldsymbol{r_{11} = 0{,}709324 \approx 0{,}71}$$

2. Data Analysis

a. Difficulty Level of Test Items

Basedon the data of item analysis result in difficulty level that the writer got as shown in table4aboveit can be knownthat from 30 items, there is 1 item equivalent to 3.33% of total test items regarded as easy test item because it is in the range from 0,00 to 0,30. It means that this item can be interpreted as not good test item because it is in difficult level.

Next, there are 13 items equivalent to 43.33% of total test item regarded asmoderate test items, with the range from 0,30 to 0,70. So, these items are moderate ones since there are many students from both upper and lower groups who answered correctly. Furthermore, 16 items equivalent to 53.33% ofthetotal number oftest items that belong to easy level of difficulty, with the range more than 0,70. It means that they are difficult items because they are answered correctly by most of students in both groups.

In conclusion,the EnglishSummativetest forclass IXat SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang Bandung at odd semester2013/2014 regarded easy test because 53.33% of total items are easy.

b. Discriminating Power

Basedon the data of item analysis result in discriminating power as shown in appendix 4 and table5aboveit can be knownthat from 30 items, most of the items are satisfactory in discriminating power because they have mediumability to differentiate between students who are in upper group and those who are in the lower group.

To sum up, discriminating power of English Summative test indicates that 3 items eqivalent to 10 % have very poor dicriminating power. Furthermore, 11 test items equvalent to 36,67 % range less than 0,20. It means that the items are poor. The rest of test items indicate 12 items equivalent to 40 % of total test items are satisfactory because they range from 0,20 to 0, 40; meanwhile 4 items equivalent 13.33 % they have good discriminating power.

In conclusion, the English Summative test for grade IX at SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang Bandung at odd semester 2013-2014 has satisfactory discriminating power because40 % of test items are satisfactory.

c. Effectiveness of Distractors

Based on the data of item analysis in efectiveness of distractor in appendixe 6 and table 6, the result shows some distractors do not distract testee well though the effectiveness of distractor should divert students who have not studied well from the correct answer. This means many students are still

not distracted by the optional answers because 56,7 % of the distractors do not work properly.

To sum up, the test has the effectiveness of distractors in the amount of 43,3% function well and 56,7% function unwell. So, the writer can interpret that the efectiveness distractors of the English Summative test for the twelfth of SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang at odd semster 2013-2014 belongs to poor level. In other words, the test has poor distrators.

d. Validityof Test Items

Basedonthe data analysis of test item validity as shown in table 7and the calculation in table 8, the researcher got the percentage of items regarded as valid and invalid items. There are 21 items equivalent to 70 % of total items said to be valid, and the rest of 9 items equivalent to 30 % of total items belong to invalid items. It means that they are valid items because the value of correlation coefficient result is greater (>) than table value (rt) = 0.213 for the 5% level, therefore the result obtained is significant. In other words, that the English Summative test items for the twelfth of SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang Bandung at odd semester 2013-2014 are regarded valid.

e. Reliability Analysis

A test to be reliable if it is consistent and dependable. This means that an elementary concern about the development and the use of English Summative test is its reliability, that is, the constancy of the test as a measure. Reliability refers to the consistency of the examination scores. Also, it refers to the scope to which the test produces consistent results if different markers mark it.

The number of correlation coefficient (r) obtained from the above data calculation is in the amount of 0.71. the correlation number of 0.71 lies between the interval 0.70 to 0.90 with a high interpretation. It can be concluded that the English Summative test for grade of SMP Negeri 02 Ngamprah and SMP Negeri 03 Padalarang at odd semester 2013/2014 has high reliability.

D. CONCLUSION

Based on the research of validity and reliability analysis on English Summative Test at private senior high school in Bandung Timur, the conclusions are as follows:

Basedon the data of item analysis result in difficulty level, It can be concluded that there are 16 items (53,33%) regarded as easy test items that range from 70 to 1,00. Meanwhile, in the discriminating power, the test items are satisfactory because 12 items equivalent to 40 % of total test items range from 0,20 to 0, 40. In the term of the efectiveness of distractor, the distractor do not function well because 17 items equivalent to 56,7 % of the distractors are ineffective.

Moreover, there are 21 items equivalent to 70 % of total items said to be valid, so the test has good validity because the value of correlation coefficient result is greater (>) than table value (rt) = 0.213 for the 5% level. In the term of reliability, the number of correlation coefficient (r) of the test is in the amount of 0.71. The correlation number of 0.71 lies between the interval 0.70 to 0.90 with a high interpretation. It can be concluded that the English Summative test has high reliability.

The writer would like to give some suggestions to addressed to test makers or the teschers as feedback of the research result. (1) The test maker should revise the items that do not belong to moderate level of difficulty or, in other words, the items that are included as easy or difficult items. (2) they should should revise the items regarded as poor or even satisfactory items in discrimination index to be good and excellent ones. (3) they should pay attention on the test distractors that still do not function well/effectively.Finally, the test maker should maintain the validity and reliability of the test and develop them through innovative and constant experimenting all the time.

E. REFERENCES

Alderson,et al.1995**.** *Language Test Construction and Evaluation*.Cam bridge: Cambridge University Press.

Alderson, Charles. (2009). *The Politics of Language Education: Individuals and Institutions***.**Clevedon: Multilingual Matters.

Allen, David. (1998). *Assessing Student Learning*. New York: Teachers College Press.

Allison, D.(1999). *Language Testing and Evaluation: an Introductory Course*. Singapore University Press: Singapore.

Al–Mashharawi, B.(2006). "*Evaluating Teachers' Performance in Teaching Speaking Communicatively in Preparatory Stage in Jabalia Area*". Unpublished Dissertation, Faculty of Education: IUG.

Al-Shumaimeri, Y. (1999). *"An Evaluation of An English Language Test".Published study*, King Saud University, College of Education: KSA.

Anderson, P. and Morgan G. (2008). *Developing Tests and Questionnaires for a National Assessment of Educational Achievement*. World Bank: Washington, USA.

Arikunto, Suharsini. 1996, *Dasar-dasar Evoluasi Pendidik'an* Jakarta: Bumi Aksara

Ary, D., Jacobs, L. C. and Razavieh, A. (1996). *Introduction to Research in Education*. New York: Harcourt Brace College Publishers.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bailey, A. (2007). *The Language Demands of School. New Haven*, Yale University Press: New Haven, USA.

Brown, H. Dougles. (2004) *language Assessment Principles and Classroom Practices*. Longman: San Francisco State University.

Brown, J. (2005) *Testing in Language Programs*. Englewood Cliffs: Prentice Hall Regents.

Brown, James Dean, *Testing in Language Programs*, New Jersey: Prentice Hall Regent, 1996.

Brown, J.D. and Hudson, T. (1998). *The alternatives in language assessment*. TESOL Quarterly 32, 653–75.

Carroll, J. B. (1968). 'The psychology of language testing' in Davies, A. (ed.) *Language Testing Symposium. A Psycholinguistic Perspective*. London: Oxford University Press. 46-69.

Choppin, B. (1990). *Evaluation, Assessment and Measurement*. In International Encyclopedia of Educational Evaluation. Edited by Walberg, H, & Haertel, G., N, Y,. Peragmon Press.

Colin Phelan, Julie Wren. (2005). *Academic Assessment; Exploring Reliability In Academic Assessment*.UNI Office of Academic Assessment

Conner, C. (1999). *Assessment in Action in the Primary School*. Falmer Press: Washington, D.C

David p. Harris (1969) *Testing English as a Second Languace*. New York: Tata Megrowhill, Inc

DesMarais, R. (2008). *Student Success Handbook.* USA: New Readers Press

Episcopal, O. (2008). *Standardized Achievement Tests at OES*.

Fulcher, Glenn and Fred Davidson. (2007) *Language Testing and Assessment.* London and New York: Routledge

Graves, K. (2000) *Designing Language Course* .Heinle & Heinle Publishers

Hahn, Nicole. (2007). *Assessing the young Learners' Progress: Tests. Muunchen*: GRIN Verlag GmbH.

Harlen, W. (2007*). Assessment of Learning*. SAGE: UK.

Heaton, J.B. 1988. *Writing English Language Test*. London: Longman

H.H. Remmers, et. al., *A practical introduction to measurement and evaluation*, (New York: Harpers and Brothers, 1960), p. 126.

Hopkins, Charles D. And Antes, L. Richard. 1990. *Classroom Measurement and Evaluation*. Cambridge: Cambridge University Press

Huitt, W. (2004, July). *Assessment, measurement, and evaluation. Educational Psychology Interactive.* Valdosta, GA: Valdosta State University. Retrieved on the 5th of May, 2010, from:http://www.edpsycinteractive.org/topics/measeval/measeval.html.

J. Stanley Ahman and Marvin D. Glock1967, *Evaluating Pupil Growth*,Principles of Test at and Measurement, third edition, Boston: Allyn and Bacon

J. Wayne Wrighstone, 1956. *Evaluation in Modern Educalion*, New York American Book Company.

Kthleen, M Bailey, *Learning about Language Assessment: Dilemmas, Decisions, and Directions*, New York: Heinle & Heinle Publishers, 1998.

Lynch, B. K. (2001). *Rethinking assessment from a critical perspective. Language Testing 18 (4)* 351–372.

Macalister (2010). *Language Curriculum Design*. New York.

Purwanto, Ngalim. 1994. *Prinsip-prinsip dan Teknik Evaluasi Pengajaran*. Bandung: PT. Remaja Rosdakarya

Richards, J. C. (2001). "Reflective Teaching in TESOL Teacher". *Issues in Language Teacher Education.* Retrieved on 13th of August 2010 from:http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED370357.

Richards, J. C. (2001). *Curriculum Development in Language Teaching*. New York: Cambridge University Press.

Robertson & Nunn. (2013).*The Study of Second Language Acquisition in the Asian Context Paperback.* The Asian EFL Journal

Sudijono, Anas, *Pengantar Evaluasi Pendidikan*, Jakarta, Raja Grafindo, 2003.

Sudjana, Nana 1991. *Penilaian Hasil Proses Belajar Mengajar*. Bandung: PT Remaja Rosdakarya

Surapranata, Sumarna 2004. *Analisis Validitas Reliabilitas dan Interpretasi Hasil Tes Implementasi Kurikulum 2004.* Jakarta: Rineka Cipta

Thorenfeldt, A. (2005). *Unpredictable and full of risks*? *An evaluation of the exam assessment in English in the R'94 vocational courses*. University of Oslonsis. Norway.

Wainer and Braun,(1988). *Test Validity:* American Education Research.

Weiss, C. H. (1972). *Evaluation Research: Methods for Assessing Program Effectiveness*. Englewood Cliffs, NJ: Prentice-Hall.

Wilmar, Tinambunan (1988).*Evaluation of Students Achievment*, Jakarta: Depdikbud

Woodfield, P. (2008). *Assessmentfor Learning and Teaching in Primary Schools Learning* Matter: UK.

Wrighstone, Wayne. (1956). *Evaluation in modern education* : New York : American Book Company.