THE RELIABILITY OF THE ENGLISH SUMMATIVE TEST FOR THE SECOND GRADE STUDENTS OF SENIOR HIGH SCHOOL

Abdul Muhid Murtadho
STIBA IEC Jakarta
muhid@stibaiecjakarta.ac.id

ABSTRACT

This research was held to check out the reliability of the English Summative test for the second grade students of the Senior High School. This Research was held at SMA 3 Tambun Selatan – Bekasi. The writer took 43 students as sample for research. The research instrument is The English Summative test of second semester for second grade students. The English Summative test is tested twice. The methodolgy used by the writer is Test – Retest and after that the writer count the correlation of those tests using formula of Pearson Product Moment. From The research writer found that the English Summative test for the second semester of second grade students in SMA 3 Tambun Selatan Bekasi period 2009 / 2010 is reliable.

Key Words : Reliability, Evaluation, English Summative Test, Pearson Product Moment

A. INTRODUCTION

1. Background

Having good education has been the priority for many people. According to W.S Winkel formal education is a kind of education which is planned and structured systematically and implemented at school (W.S Winkel, 1983:159). School is a place where students can learn and study. In Indonesia formal education is categorized based on the ages and the capability of the learners. It consists of some levels, they are Elementary school, Junior high school, and Senior high school.

Teaching and learning only are not enough. They are not completed yet, Tony Wright says that we must always bear in mind that teaching / learning activity is a long term process, and the changes of participant's behavior and knowledge are usually difficult to evaluate and measure (W.S Winkel, 1983:159). We can not only teach the students without knowing how much knowledge they have achieved. In this case the school and teacher need to measure the capability of the students or learners during the learning process by evaluating the students. In formal education, it is known as a learning evaluation. Evaluation is used to know and measure whether the objectives have been obtained or not. It is an answer of the question "How good the students are".

There must be a good and bad test in school / institution. It is obviously that an appropriate test should have good criteria as follows; validity, reliability, and practicality. As stated above, one of the criteria of a good test is reliability. The indication that a test is reliable is the test can give consistency or stable result of the same pupils in any occasion. A test can be ascertained reliable if a test can give true score. On the contrary, a test can't be ascertained reliable if the test does not give stable result of the same value of the same pupils in any occasion. It is also stated by John A.S Read, "Reliability refers to the accuracy, consistency, and stability of measurement by a test" (John A S Read, 1981:4). Otherwise, there are many kinds of test. According to Michael Harris and Paul McCan, test can be classified in terms of their functional role in classroom instruction. They are: Progress test, Placement test, Proficiency test, Diagnostic test, and Summative test (Michael Harris and Paul McCann, 1994:28).

The summative test is used to determine terminal performance and this test typically given at the end of the instruction. For measuring the teaching and learning activities which is done at least four months every semester, SMA 3 Tambun Selatan – Bekasi gives a test called summative test for their pupils.

Based on the information that the writer obtain from the competent resource; the second year class teacher at this school, the quality especially the reliability of the English summative test in the first semester of the second year class of SMA 3 Tambun Selatan - Bekasi has never been known yet. Even though the test is made by a professional team consisting of well-experienced educators or teachers, the summative test cannot be ascertained reliable yet. It is possible that the test is not reliable because no analysis has been conducted to get empirical evidence in determining the reliability of it. This becomes the reason why the writer is interested in conducting the research to find out whether the English summative test of the second semester of the second year class at SMA 3 Tambun Selatan - Bekasi academic 2009 / 2010 is reliable or not reliable.

2. Objectives

The writer main purpose in conducting the research is to get empirical evidence as the answer of the problem statement. It is to find out whether the English summative test in the second semester of the second year class of XI IPA 1 in SMA 3 Tambun Selatan - Bekasi of the academic year 2009/ 2010 is reliable or not.

3. Theoretical Framework.

1) Teaching and Learning

Teaching and learning are two things that can not be separated. Both of them are done by different persons, teaching is conducted by a teacher and learning is accomplished by a learner. Nowadays, teaching and learning are always noticed in every institution or school. Many language experts state about the definition of teaching and learning. Most of them have similar opinions and perception about it. According to Brown teaching may be defined as “showing or helping someone to learn how to do something, giving instruction, guiding in the study or something, providing with knowledge, causing to know or understand” (H.Dougles Brown, 2007:8). The point of that definition is helping and guiding. Teachers must be able to help and guide the learners to understand the lesson or materials. One way to help and guide the learners to obtain the lesson easily is by setting the exciting and enjoyable situation in class.

The definition of learning stated by Dressel and Marcus learning is a process in which the learner attends to surrounding circumstances and is changed by exposure to them (Paul L Dressel and Dora Marcus). The key word from that statement is process and change. Learners will get progress and change when they learn seriously and continually. Learners always expect the changes and progress.

Based on the statements above the writer summarizes the concept of teaching andlearning. Both are as processes. Teaching is process to transfer the knowledge which is given by a teacher and learning is the process to get the knowledge which is done by students.

2) Evaluation

Evaluation is an integrated activity which cannot be separated from education and classroom activities. It is an intrinsic part of teaching and learning. It is very important for the educational system and for the teacher. Purwanto says that, *“One of the most effective way to change the teaching process is by evaluating the result of the test which is gotten from the teaching and learning process it self*” (M.Ngalim Purwanto, 2001:118) The improvement and the change are about the improvement of the students’ quality.

The writer summarizes that evaluation is an intrinsic part of teaching and learning activity, which can provide valuable information especially for the teachers in making the educational decisions for the future direction of classroom practice and for the planning and management of learning tasks and students.

3) Test

There are some words related to test. They are testing, tester, and testee. Here are the explanation according to Annas Sudijono, Test is the instrument or procedure which used for measuring and scoring the students. Testing is the time when we do the test. Tester is the person who makes experiment about the test or the test maker and testee is the person who does the test or the doer or the test participants (Anas Sudijono:66).

As written above the test is an instrument, it is briefly stated by the experts. Tuckmen says that tests are tools that are useful in a number of process such as evaluation, diagnosis, or monitoring (Bruce W Tuckmen, 1975:12). The key point of that statement is tool. Tools are synonym of the instruments, so we could say tests are the instruments or tools used to evaluate the students. Test not only benefit the teacher but test also can benefit students or even administrator by confirming progress that has been made and showing how we can best redirect our future effort.

The writer resumes that test is a mean or device that can be used by a teacher to measure and evaluate the capability of their students as the answer of “How well” they are. A test is also a procedure that is used to assess the testee and the overall efficiency of improving teaching and learning activity. Test can be formed as essay or multiple choices. Test also provides comparability of students’ performance for a certain time.

4) **Type of Test**

There are some kinds of test based on their functional role in classroom instruction. According to Gronlund there are four kinds of test, they are Placement test, Formative test, Diagnostic test, and Summative test (Norman E Gronlund, 1981:17).

Summative test is evaluation of pupil achievement at the end of instruction and the function is to evaluate achievement at the end of instruction. Here are some definitions by the experts. Grounlund defined the summative as the achievement test that is given at the end of period of instruction for the purpose of certifying or assessing grades (Norman E Gronlund, 1981:5). The key word from that definition is at the end and certifying grades. The end means the time when the teaching and learning activities have finished. Certifying grades means checking or evaluating the capability of the learners whether they have achieved the goals or not.

The writer summarizes that summative test is a kind of test which is given to evaluate the students or learners during the learning process in every semester and it is usually held at the end of instruction. Teachers can know how well their students are and know whether they can continue to the higher level. It is also used to assess students’ achievement.

5) The Characteristic of Good Test

According to some experts, characteristics of good test can be defined in some ways. Lado said that, the question we ask about a test will vary in each case depending on purpose, time subject, etc. in general, however, we must ask if a test is valid, reliable, scorable, economical, and administrable (Robert.Lado,Ph.D., 1961:30. That is clear that we must make our test valid, reliable, scorable, economical, and administrable. Harrison states, "The most important characteristics of a good test are reliability, validity, and practicality (Andrew Harrison.1983:10). From both of them, it is clear that validity, reliability, and practicality are the most important characteristic of good test.

(a) Reliability

Test needs to be reliable to know whether the test can be trusted according to the criteria which have been formulated. A test can be ascertained, if a test can give same or similar score when it is given to the same respondent in different occasion. Harrison said that reliability of a test is its consistency (Andrew Harrison:10). It means that reliability refers to consistency measurement. It is to know how consistent test scores or other evaluation results are from one measurement to another.

To designate a test's accuracy, the term reliability is used. It is same when we ask about the test's reliability, we are not asking what it measure, but instead how accurately it measures whatever it does measure. Allison said that, the reliability of a test concerns the accuracy and trustworthiness of its result (Desmon Allison:85).

(b) The Method of Estimating Reliability.

There are three widely methods for assessing reliability according to Lado, they are Retesting method (Test – retest method), Equivalent forms method (Alternate-form method), and Split half method (Chance Half method). Below are the definition and the explanation about the methods (Robert Lado,Ph.D :332).

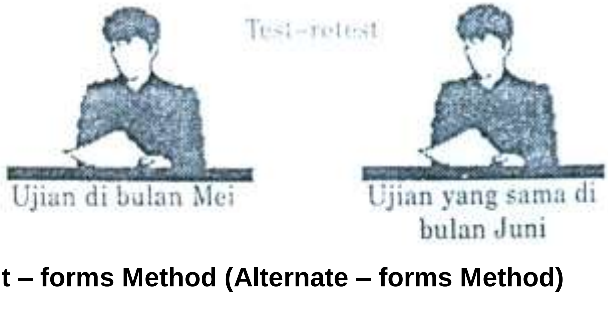
6) Test - Retests Method. (Retesting Method)

The test re-test method is one of ways in estimating the reliability of a test by re

administrating the test. In this method, one set of test is administered twice to the same student and compute the correlation between two sets of scores. According to Gronlund, *"To estimate reliability by this means of the test re-test method, the same test is administered twice. The resulting test scores are correlated, and this correlation coefficient provides a measure of stability; that is, it indicates how stable the test results are over the given period of time. If the results are highly stable, those pupils who are high on one administration of the test will tend to be high on the other administration, and the remaining pupils will tend to stay in their same relative positions on both administrations" (*Norman E Gronlund.1985:90)*.*

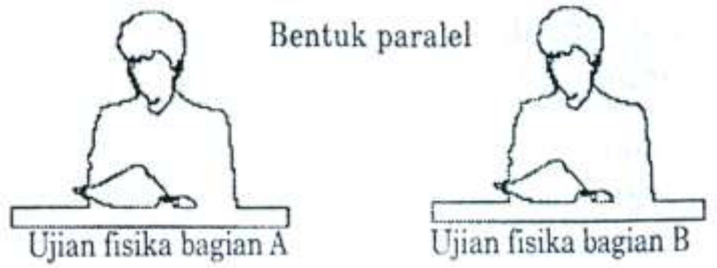
From that statement it clear that test- retest can be given about one or two weeks after the first test given. Test – retest is one of the three basic reliability strategies; this method is the most appropriate for estimating the stability of a test over time. It is briefly stated by Brown, test – retest strategy is the most logical and practical alternative for estimating reliability (James Dean Brown:193). It is clear that test – retest is the most appropriate method for estimating the stability. Surapranata also stated about it. He said that *test – retest method is very useful to see the stability of measurement (*Sumarna Surapranata:97). It means that test- retest method is important and useful to estimate the reliability of one test.

Below is the figure of estimating reliability with test-retest method according to Sumarna Suraprnata *(*Sumarna Surapranata:97).



b. Equivalent – forms Method (Alternate – forms Method)

Another method that can be used to estimate the reliability is equivalents – forms methods.



c. Split – Half Method (Chance Half Method)

Another way to estimate reliability is split-half method.

From three methods, It is clear that test – retest is the most appropriate method for estimating reliability as mentioned by Brwon and Surapranata, based on that reason the writer is going to use the test – retest method to measure the stability and the equivalence of the reliability coefficient of the English summative test of the second semester for the second year class at SMA 3 Tambun Selatan of the 2009-2010 academic year.

4. Methodology

This research uses an evaluation research, Evaluation research is the systematic process of collecting and analyzing data about the quality, effectiveness, merit, or value of programs, products and practices (L.R Gay, et all. 2006:7), which implements the correlation analysis of the two groups of scores: the English summative test of second semester in SMA 3 Tambun

Selatan – Bekasi period 2009 / 2010 scores as variable X, and the retest of English summative test in SMA 3 Tambun Selatan – Bekasi period 2009 / 2010 scores as variable Y.

The data is analyzed by estimating the validity and the reliability. It is stated by Sugiyono, *"The reliable instrument may not be valid, but the valid instrument commonly must be reliable*" (Sugiyono, 2002:268). The point of that statement is that validity always related to the reliability, an instrument will be reliable if that instrument is valid and in opposite if the instrument is not valid, that instrument will not be reliable.

In addition, to improve the process of teaching and learning at school we need to evaluate the teaching and learning process by two ways, the first is by making the item analysis and the second one is by counting the validity and the reliability of the instrument. That opinion is stated by Ngalim Purwanto.

He said that, *"The process of the result for the test in changing the process of teaching and learning can be done by two ways:* first By Making the analysis of items and the second By Counting the validity and the reliability of the test (M.Ngalim Purwanto).

Based on those reasons the writer is going to analyze the English Summative test of second semester for second grade at SMA 3 Tambun Selatan – Bekasi period 2009 / 2010 by counting the validity and the reliability. Then after the validity and reliability are known, and the instrument has been valid and reliable, the writer will make an item analysis to emphasize the result of the data. The purpose of making item analysis is to know whether the test is good or need revision. Purwanto explains about the purpose of making item analysis. Item analysis is used to find out which test is good and bad and why the item of the test can be good or bad (M.Ngalim Purwanto).

To analyze the validity of this instrument the writer uses the *pearson product moment.* This formula is suggested by Burhan Nurgiyantoro, and to analyze the reliability of this instrument the writer also use the *person product moment* as suggested by Burhan Nurgiyantoro (224:238). However to analyze item analysis the writer is going to use the Item difficulty level or item facility and the formula of discriminating power or item discriminabilty as suggested by Burhan Nurgiyantoro (1987:126).

Below is the explanation of the ways and steps about how to analyze the instrument by calculating the validity, reliability, and item analyzes.

1) How to analyze the validity of the instrument

a) Make the work table which consists of Variable X, Variable Y, Variable X ^{2,} , Variable Y ^{2} , and Variable XY

b) Count and find the $\sum$ of each rows.

c) Put the numbers on the pattern suggested.

$$r = \frac{N\sum X_1X_2 - \left(\sum X_1\right)\left(\sum X_2\right)}{\sqrt{N\sum X_1{}^2 - \left(\sum X_1\right)^2 N\sum X_2{}^2 - \left(\sum X_2\right)^2}}$$

d) After the (r) is known, we need to consult a table. (see appendix 4)

e) The (r) table will tell us whether the score represent true relationship or not by using the pattern suggested by Burhan Nurgiyantoro,*“ Jika koefesien korelasi* (r) *yang diperoleh* $\geq$ *daripada koefesien di table nilai - nilai kritis* (r) *table, yaitu pada taraf signifikansi 5% atau 1% , instrument tes yang diujicobakan tersebut dapat dinyatakan valid.”* or it can be seen below:

$$(r) \geq 5\% \geq 1\%$$

2) How to analyze the reliability of the instrument

a. Make the work table which consist of Variable X, Variable Y, Variable X ^{2,} , Variable Y ^{2} , and Variable XY

b. Count and find the $\sum$ of each rows

c. Put the numbers on the pattern suggested

$$r = \frac{\sum XY}{\sqrt{\left(\sum X^2\right)\left(\sum Y^2\right)}}$$

d. After the (r) is known, we need to consult a table. (see appendix 4)

e. The (r) table will tell us whether the score represent true relationship or not by using the pattern suggested by Burhan Nurgiyantoro,

$$(r) \geq 5\% \geq 1\%$$

3) How to analyze the item analysis.

a. Arrange the score of the students answer sheet form the highest to the lowest.

b. Take 27,5% from the highest score and 27,5% from the lowest score. The first group can be named ***Upper group*** and the second group can be named ***Lower group***. The rest of the sample can be named ***Middle group.*** (see below)

{ } 27,5% = Upper group

{ } Middle group

{ } 27,5% = Lower group

c. Analyze the correct and the wrong answer for each item. This analyses just the upper group and lower group, and we just ignore the middle group as well. Based on the analyses we will get the index for the item difficulty and the item discriminabilty.

4) The formula of Item difficulty and Item Discriminabilty

a. Item difficulty sometimes called item facility, it is the statement about how easy or difficult is the test items for the students.

Below is the Formula of Difficulty level or Item Facility:

$$IF = \frac{FH + FL}{N}$$

index of item facility is between **0,15 – 0,85**

b. Formula of Discriminating power or Item Discriminabilty

Item discriminabilty means how good is the test item can differentiate

between the higher group and the lower group. Burhan said that, " *Butir soal yang*

baik adalah yang dapat membedakan antara dua kelompok secara layak" Below is the formula of item discriminabilty

$$ID = \frac{FH - FL}{n}$$

index of item discriminabilty is between **-1,00 – 1,00** or it means the index of item discriminbilty minimum is **0,25**

B. RESULT AND DISCUSSION

1. Research Findings

1) Test Composition

The test consists of 40 numbers. Number 1 until 40 is multiple choices. The multiple choices consist of reading, structure, and vocabulary test. The writer is going to analyze all the multiple choices as the research.

2) Description of Data

The data are taken in SMA 3 Tambun Selatan – Bekasi class XI IPA 1 that consist of

43 students. The first test was held on June 15th, 2010 and this first test is used as the variable X. The second test or the retest test was held on June 21st, 2010 and this second test is used as the variable Y.

3) Analysis of Validity

To analyze the validity of the English summative test of the second semester of the second grade in SMA 3 Tambun selatan – Bekasi period 2009 /2010, the writer uses the formula as follow:

$$r = \frac{N\sum X_1X_2 - \left(\sum X_1\right)\left(\sum X_2\right)}{\sqrt{N\sum {X_1}^2 - \left(\sum X_1\right)^2 N\sum {X_2}^2 - \left(\sum X_2\right)^2}}$$

The formula above is suggested by Burhan Nurgiyanto. The instruments of the test can be valid if the values of coefficient correlation (r) are ≥ than coefficient in table score of (r) product moment on appendix. The significance is about 5% or 1%. Here is the analysis.

It is known that :

N	=	43
$\sum X_1$	=	1034
$\sum {X_1}^2$	=	25388
$\sum X_2$	=	1041
$\sum {X_2}^2$	=	25627
X_1X_2	=	25384

Now we can put them as the pattern given:

$$r = \frac{43(25384) - (1034).(1041)}{\sqrt{\{43(25388) - (1034)^2\}\{43(25627) - (1041)^2\}}}$$

$$r=\frac{1091512-1076394}{\sqrt{(1091684)-(1069156)(1101961).-(1083681)}}$$

$$r=\frac{15118}{\sqrt{(22528).(18200)}}$$

$$r=\frac{15118}{\sqrt{410009600}}$$

$$r=\frac{15118}{20248{,}694}$$

$$r=0{,}746$$

To know the significance correlation coefficient (r) we need to see the value score tables (r) product moment. (See appendix 4). For The first we need to decide *"Degree of freedom"* (df), as suggested by L.R Gay, " Number of participants affects the degree of freedom, which for the Pearson r are always computed by the formula $(N)-2$". Or it can be seen like this $df=n-2$ (L.R Gay et al:329). So, to find (df) we can do like this, $df=43-2=41$. Based on the value score table on the appendix 4, the significance for each 5% and 1% are 0,308 and 0.398. Then we use the pattern that suggested by Burhan Nurgiyantoro, it can be seen like this:

$$(r)\geq 5\%\geq 1\%$$

$$0{,}746\geq 0{,}308\geq 0{,}398$$

From the result above, it is clear that the coefficient correlation is higher than the (r) table from 5% or 1%. So, it means the instrument can be said **valid**.

4. Analysis of Reliability.

1) Description of Variable X (The first Test of English Summative test)

Based on the first test of English Summative test score on appendix 3, the writer obtained $\sum X = 1034$, $\sum X^2$ = 25388 with 43 respondents.

2) Description of Variable Y (The second Test of English Summative test)

Based on the second test of the English summative test on the appendix 3, the writer obtained $\sum Y = 1041$, and $\sum Y^2 = 25627$ with 43 respondents.

Burhan Nurgiyantoro suggested to analyze the reliability of instrument we can use the formula *pearson product moment*. Here is the pattern:

$$r = \frac{\sum XY}{\sqrt{\left(\sum X^2\right)\left(\sum Y^2\right)}}$$

Above we have gotten the score of $\sum X^2 = 25388$ and $\sum Y^2 = 25627$. We also have gotten the score of $\sum XY$ on the Appendix. $\sum XY$ = 25384. so now we can put the score on the pattern. Here they are:

$$r = \frac{25384}{\sqrt{(25388)(25627)}}$$

$$r = \frac{25384}{\sqrt{650618276}}$$

$$r = \frac{25384}{25507{,}22}$$

$$r = 0{,}996$$

To know the significance correlation coefficient (r) we need to see the value score tables (r) product moment. (See appendix 4). And then we need to decide *"Degree of freedom"* (df), as suggested by L.R Gay, " Number of participants affects the degree of freedom, which for the Pearson r are always computed by the formula $(N)-2$". Or it can be seen like this $df = n-2$. So, to find (df) we can do like this, $df = 43-2 = 41$. Based on the value score table on the appendix 4, the significance for each 5% and 1% are 0,308 and 0.398. Then we use the pattern that suggested by Burhan Nurgiyantoro, it can be seen like this:

$$(r) \geq 5\% \geq 1\%$$

$$0{,}996 \geq 0{,}308 \geq 0{,}398$$

From the result above, it is clear that the coefficient correlation is higher than the (r) table from 5% or 1%. So, it means the instrument can be said **reliable**.

Based on the answer of the analysis of the validity and reliability, and the statistical hypothesis on page 54, it can be said like this, Null hypothesis (**Ho**) is rejected and Alternative hypothesis (**Ha)** is accepted, it means there is correlation between variable X and Y.

5. Analysis of Items

Below is the item analysis of the English summative test of second semester for second grade at SMA 3 Tambun Selatan – Bekasi. The item analysis will be counted using the item facility and item discriminability. Here they are:

1) Item Facility

Below are the ways of counting the items facility based on the formula:

$$IF = \frac{FH + FL}{N}$$

Sample item number 1 : $IF = \frac{11+4}{24}$ = 0,62

Sample item number 2 : $IF = \frac{3+0}{24}$ = 0, 12

From the result above we can say that:

Item number 1 = **0, 62** it is between **0,15 – 0,85** it means this item is *Accepted*

Item number 2 = **0, 12** it is not between **0,15 – 0,85** it means this item is *Rejected*

2) Item discriminability

Below are the ways of counting the items facility based on the formula:

$$ID = \frac{FH - FL}{n}$$

Sample item number 1 : $ID = \frac{11-4}{12}$ = 0,58

Sample item number 2 : $ID = \frac{3-0}{12}$ = 0, 25

From the result above we can say that:

Item number 1 = **0,58** The minimum score is **0, 25** it means this item is *Accepted*

Item number 2 = **0,25** The minimum score is **0, 25** it means this item is *Rejected*

From the result of the item facility and item discriminability, the writer concludes each item as follow:

Item number 1 : IF = **0,62** and ID = **0, 58** it means this item is ***Accepted***

Item number 2 : IF = **0, 12** and ID = **0, 25** it means this item is ***Rejected***

These are list of items number that has been reasonable and need to be revised.

Items which are accepted	**Items which need to be revised**
1, 9, 12, 13, 15, 16, 17, 26, 27, 29, 30, 31, 32, 38, 40	2, 3, 4, 5, 6, 7, 8, 10, 11, 14, 18, 19, 20, 21, 22, 23, 24, 25, 28, 33, 34, 35, 36, 37, 39

C. CONCLUSION

Based on the result of the English summative test, the research shows that there is positive and significance correlation between the first English test set (data X) and the second test or retest of the English test set (data Y) of English Summative test for the students of second grade at the second semester of SMA 3 Tambun Selatan – Bekasi period 2009 / 2010. It means the Alternative hypothesis (H_1) is accepted and the null hypothesis $(H_0)$is rejected. Overall, the writer concludes that the English Summative test for the second grade at the second semester of SMA 3 Tambun Selatan – Bekasi period 2009 /2010 is ***reliable***.

BIBLIOGRAPHY

Ahman, J Stanley and Marvin D Glock.1967.*Evaluating Pupil Growth Principal Tests and Measurement*.United States of America.Allyn and Bacon.

Allison, Desmon. 1999. *Language Testing and Evaluation.*Singapore. Singapore University Press.

Brown, H.Dougles. 2007. *Principles of Language Learning and Teaching*. United States of Amaerica. Pearson Education

Crowl, Thomas.K., 1996, *Fundamentals of Educational Research,* United States of America, Brown and Benchmark Publishers.

Dickins, Paulina Rea- and Kevin Germaine.1992.*Evaluation*.London. Oxford University Press.

Dressel, Paul L and Dora Marcus. 1982. *On Teaching and Learning in College*. United States of America. Jossey – Bass Publishers.

Gay, L.R, et al. 2006. *Educational Research.* Pearson : Merrill Practice Hall

Genesee, Fred and John A Upshur.1996.*Classroom-Based Evaluation in Second Language Education*.Australia.Cambridge University Press.

Gronlund, Norman E.1981.*Measurement and Evaluation in Teaching*. United States of America. McMillan Publishing Co.Inc

Hamalik, Oemar. 2001. *Proses Belajar Mengajar.* Jakarta. Bumi Aksara.

Harmer, Jeremy. 2007. *How to Teach English*. England. Pearson Education Limited

Harris, Michael and Paul McCann. 1994. *Assessment*. Scotland. Heinemenn English Language Teaching.

Harrison, Andrew.1983. *A Language Testing Handbook*. London. The McMillan Press Limited.

Hopkins, Charles D and Richard L Antes. 1990. *Classroom Measurement and Evaluation*. Itasca. Peacock Publisher.

Lado, Robert,Ph.D.1961.*Language Testing*.United States of America. McGrow-Hill Book Company.

Nurgiyantoro, Burhan et al, 2004, *Statistik Terapan*. Bandung, Gajah Mada University Press.

Nurgiyantoro, Burhan, 1987, *Penilaian dalam Pengajaran Bahasa dan sastra*, Yogyakarta, BFPE

Oller, John W.jr. 1979. *Language Tests at School*. London. Longman Group Limited

Purwanto, M Naglim, 2001, *Prinsisp – prinsip Evaluasi Pengajaran* , Bandung,

PT.Remaja Rosdakarya.

Read, John A S. 1981. *Papers on language Testing*. Singapore. Seameo Regional Centre.

Sudijono, Anas.1996.*Pengantar Evaluasi Pendidkan*. Jakarta. Rajagrafindo Persada.

Sugiyono, 2002, *Statiska Untuk Penelitian*, Bandung CV.ALFABETA

Surapranata, Sumarna.2004.*Analisis, Validitas, Reliabilitas, dan Interpretasi Hasil Test Implementasi Kurrikulum 2004*.Bandung.PT.Remaja RosdaKarya

Thorndike, Robert M.et.al.1991.*Measurement and Evaluation in Psychology and Education*.Newyork.McMillan Publishing Company

Tuckmen, Bruce W. 1975. *Measuring Educational Outcomes Fundamentals of Testing*. United States of America. Harcourt Bruce Jovanovich.

Walkin, L. 1982. *Instructional Techniques and Practice*. United Kingdom. Stanley Thornes

Weir, Cyril. 1993. *Understanding and Developing Language Test*. UK. Prentice Hall International

Winkel, W.S.1983.*Psikologi Pendidikan dan Evaluasi Belajar.* Jakarta. PT.Gramedia.

Wright, Tony 1987. *Roles of Teachers and Learners*. London. Oxford Univrsity Press.

Wrightstone, J. Wayne.1956.*Evaluation in Modern Education*. New York: American Book Company.