QUALITY ANALYSIS OF MULTIPLE CHOICE TEST AND CLASICAL TEST AT X GRADE STUDENTS OF SENIOR HIGH SCHOOL

Trio Putra Siregar[1], Edy Surya[2], Edi Syahputra[3]

[1]*College student, Graduate Program School in Mathematics Education, State University of Medan, Indonesia*
[2]*Lecturer, Graduate Program School in Mathematics Education, State University of Medan, Indonesia*
[3]*Head, Graduate Program School in Mathematics Education, State University of Medan, Indonesia*

ABSTRACT

This study aimed at investigating the quality of multiple-choice items test created by teachers of mathematics on the topic of logical mathematic for tenth graders at Pangeran Antasari Senior High School in Medan. The test consisted of 40 multiple-choice items, and the classical theory analysis was carried out. The results turned out that 22 out of 40 items were valid and 18 items were not. The coefficient reliability resulted in 0,88196 meaning that it had higher reliability and consistency over time. Furthermore, the item-difficulty test revealed that 17 items were easy, 4 items were moderate, and 1 item was difficult. The distractor analysis indicated that 21 items were good and 1 item was very good. Of 22 valid items, there were good and tricky items existed. It was reasonable to conclude that 22 out of 40 items created by the mathematic teachers were categorized as good.

Keywords: *Test quality, Classical theory*

Preliminary

Bruce, Weil and Calhoun (Sumiati and Asra: 2008) states that learning is essentially a complex process (complex), but with the same purpose of providing learning experiences to students in accordance with the objectives. The goal is actually a reference in the implementation of the learning process. To determine the achievement of learning objectives it is necessary to evaluate learning outcomes.

According to Tyler (Rashid and Mansour: 2008), the evaluation is the process of determining the extent to which educational goals have been achieved. Broader definition put forward by the other two experts, the Cronbach and Stufflebeam (2000), the additional definition is that the evaluation process is not simply measure the extent to which the objectives are achieved, but it is used to make decisions. One way to evaluate learning outcomes is to use the test results of learning. In order to learn the test results can be used as its function is to measure the achievement of learning objectives, one of the teacher's task is to evaluate the device tests that have been made, such as with the test item analysis to determine the quality of the tests that have been made. But in reality, not many are doing so. Event analyze the test item is an activity that must be done to improve the quality of teachers that have been written test.

Darwyan Shah et al. (Arifin: 2009) defines the test item analysis as an investigation or a study of a part of the whole thing must be answered by learners. Nana Sudjana (2009) define that test item analysis or item analysis is assessment test questions in order to obtain a device that has the question of adequate quality.

From the definition above can be concluded that the analysis of items that is a process that is carried out to investigate, researching and reviewing the test questions in order to obtain a device that has the question of adequate quality. There are several reasons why the analysis of test items required. According to (Asmawi Zainul, et al: 1997) these reasons, among others:

a. To know the strengths and weaknesses of test items, so do the selection and revision of items.
b. To provide information on the specifications of items in full, so that will make it easier for device makers in formulating questions about the exam that will meet the needs in the field and a certain degree.

c. To quickly be able to know the issues contained in items, such as: ambiguity items, an error put the key answers, questions that are too difficult and too easy, or matters that have a different power is low. This problem is known immediately if it is possible for the manufacturer to make a decision about whether the items in question will be disqualified or revised in order to determine the value of learners.

d. To be used as a tool to assess the items that will be stored in a collection of matter.

e. To obtain information about the items making it possible to draw up some questions that parallel devices. The preparation of such a device is very useful when going to conduct re-examination or measures the ability of some groups of test takers in a different time.

A good test to be valid and reliable. In the view of Samuel Messick, validity of a thorough assessment which empirical evidence and logic theory to support decisions and actions based on test scores or models of another assessment (Messick: 1989). The validity of a test can be performed in various forms such as *content validity*, *criterion validity* and *construct-related validity*. Although ideally validation can be done by using all forms of the validity of such tests, but the test developer can choose a form of validation by looking at test development purposes. Kumaidi (1994) say apart from a valid, good measuring tool must also be reliable. In view of Aiken a test said to be reliable if the scores obtained by participants are relatively the same despite repeated measurements. Aiken, LR (1987) To obtain the same score, then there should be no measurement error. Thus, the reliability of a measuring instrument can be seen from the two instructions are standard error of measurement and reliability coefficient. Both of these statistics each have advantages and limitations. Feldt, LS & Brennan, RL (1989) said that in addition to a valid and reliable test that is good also depends on the number of test items contained in the good category test. More and more items are good, the better the test device. Conversely, more slightly the amount of good test, more dilapidated that test. To see the quality of a test can be conducted by using qualitative analysis (theoretical) and quantitative (empirical). In the qualitative test is said to be good if it meets the requirements of the preparation of the material, construction and language. As it kuantiatif can be done either by classical test theory techniques *(classical true-score theory)*.

Classical Test Theory

One of the world's oldest measurement theory that *behavioral* measurement are *classical true score theory*. This theory in the bahasa is often called the classical test theory. Classical theory test is a theory that is easy in its application and models are quite useful in describing how the errors in measurement can affect a score of observation. From these assumptions are then translated into several conclusions. There are seven kinds of assumptions in this classical test theory. Allen & Yen outlines assumptions theory classical as follows (Allen, MJ, & Yen, WM: 1979).

1. The first assumption classical test theory is that there is a relationship between the scores appear (observed score) are denoted by the letter X, the true score are denoted by T and *error* score which is denoted by E. According to Saifuddin Azwar (2001: 30) referred to the measurement error in the classical theory is a deviation appears of scores theoretical expectation that occurs randomly. That relationship is that big score seemed to be determined by the score pure and measurement error. In. the language of mathematics can denoted by X = T + E.
2. The second assumption is that the pure score (T) is the expected value. Therefore pure score is the average value of the acquisition of theoretical score if it were done measurements repeatedly (to infinity) against someone using a measuring instrument.
3. The third assumption classical test theory states that there is no correlation between the score of the mummy and score measurements in a test carried out. The implication of the assumption is that the pure high score will not have *an error* which is always positive or always negative.
4. The fourth assumption declare that the correlation between errors in measurement errors in the first and second measurement is zero. This means that scores of errors on two tests to measure the same thing have no correlation (relationship). Thus the magnitude of the error on a test does not depend on another test errors.
5. The fifth assumption states that if there are two tests to measure the same attributes then the error score on the first test did not correlate with pure score on the second test of this assumption would fall if one of these tests it turned measure aspects affecting the measurement error on the other.
6. The sixth assumption of classical test theory is serving on the definition of parallel test. Two sets of tests can be considered as parallel tests if scores populations taking the two tests scored the same pure and variants scores the same mistake. In practice, this theory difficult sixth assumption is met.
7. The final assumption of classical test theory states on the definition of tests equivalent (τ *Essentially equivalent).* If two devices have the test scores acquisition *X t* 1 and *t* 2 *X* which satisfy the assumptions of 1 to 5, and if for any subject population X1 = X2 + C12, where C12 is a constant se fruit number, then the second test is called the parallel test. Assumptions of the classical theory as

mentioned above allows for was developed in order to develop various formulas that are useful in measuring psychological. Different power, difficulty index, distractor effectiveness, reliability and validity are important formula derived from classical test theory.

Distinguish Power

Distinguish Power of test item is an items' ability to discriminate between higher student ability and lower student ability. Distinguish power can be determined by higher or lower of distinguish index or a number that indicates the size of the distinguish power. The function of the distinguishing features are detected individual differences are smallest among the participants of the test. Pinpointing the different grains usually done using correlation index, discrimination, and the alignment index item. Of the three ways are the most commonly used is the correlation index. There are four kinds of correlation technique which is used to calculate the different power, namely: (1) technical *point biserial,* (2) *biserial* techniques, (3) phi techniques, and (4) tetrachorik techniques.

Brennan (1972) as cited Yen WM in the Encyclopedia of Educational *Research* introduces how to calculate the discrimination index by using the following formula (Yen, WM (1992).

$$B = \frac{U}{n_1} - \frac{L}{n_2}$$

Where from the above formula can be interpreted that the distinguish power is the difference between the proportion of the group who answered correctly on the test item $\frac{U}{n_1}$ with the proportion of the group who answered the correct grain under test $\frac{L}{n_2}$,. The formula can be used to calculate the different power point items in the form of multiple choices.

Difficulty Index

Difficulty index of test item as stated by Allen & Yen is the *proportion of examinees who get that item correct.* In line with them, Sax wrote that the index of difficulty is the proportion of examinees who answered correctly Sax, G. (1980). Saifuddin Azwar (2003) states more succinct that item difficulty index is the ratio of grain answering correctly and answering many grains. The proportion of correct answer *p (proportion correct)* is the index of difficulty matter most simple and often used in determining the amount of the index. The formula for determining the magnitude of difficulty index is mathematically formulated by Saifuddin as follows:

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

P is the item difficulty index, n i is the number of test takers who answered correctly and N is the number of students who answered the item was. Thus, to calculate the index item difficulty do not split the group with the test participants into groups of top and bottom as well as to determine the difference. The magnitude of the correlation index ranges between 0 and 1. The higher the magnitude of the correlation index of the item was easier. And the smaller numbers then the correlation index items the increasingly difficult.

Difficulty index that was around 0.5 are considered the best. Because of this, according to Allen & Yen good difficulty level is 0.3 to 0,7.19 Item difficulty level below 0.3 are considered items which are difficult while if the index is above 0.7, the item was considered easy. From the above there are some things that can be inferred with regard to difficulty index point is that the *p-value* for an item only shows the index for the group tested. Price *p* this could change if the test is tested in different groups. In addition, the index of difficulty resulting from this formula is applicable difficulty index for the group. as a whole rather than the individual. Difficulty index for each participant tests can't be inferred by looking at the proportion of the index to answer correctly *p.*

Effectiveness Distractors

Each multiple-choice tests have one question and several answer options. Among the selection of answers, only one is correct. In addition to the correct answer, is the wrong answer. Wrong answer that is known as the *distractor* (detractors). Thus, distractors effectiveness is how well the wrong choice can outwit the test participants who did not know the answer key provided. The more participants who chose distractor tests, then distaktor it can function properly. How to analyze the functions distractor can be done by analyzing the pattern of spread of answers grains. The pattern of the spread of answers as stated Sudijono is a pattern that can describe how the test taker can determine the choice of the answer to the possibilities of answers that have been attached to each item (Anas Sudijono: 2005). According to Fernandes (1984) distractor said to be good if selected by at least 2% of all participants. Distractors do not meet these criteria should be replaced with other distractors which may attract more participants test to select it. Although the use of classical test theory is relatively easy to analyze the grain, but this theory has some fundamental flaws. The main drawback of classical test theory as disclosed Sumadi Suryabrata is attachment to the theory of measuring instruments in the sample *(sample bound).*

Sumadi Suryabrata. (2004) The ability of a group of students who take the test greatly affect the statistics. so the value of the statistics would be different if the test is given to another group. In addition, estimates of the ability of participants depends on items. If the index of difficulty is low then estimate a person's ability to be high and vice versa. Measurement error estimates do not include the individual but the group together. This is because each participant's response to the test question can not be explained by classical test theory. In the process of learning these things will cause various hardships, especially to see the ability of the test participants individually. Hence there is an attempt to free the measuring devices of the attachment of the sample *(sample-free).* Departing from this that the experts then draw up new theories that are intended to supplement and improve the weaknesses that exist in the classical test theory. This theory became known as *Item Response Theory* (IRT) or item response theory.

Validity Test Item

The validity test is degree validity or accuracy valid test item is the test that measure what it want to measured. So, validity test showed the level of accuracy test in measuring the target to be measure.

The validity of the test were statistically analyzed by the type of data collected. Discrete data (eg, objective test results) is calculated by the correlation point biserial. Whereas continuous data (eg test results description or attitude scales) used Pearson product moment correlation. In this paper to test the validity of the test used formula biserial correlation point for data collected in the form dichotomy (0.1) with the testing criteria $r_{tabel} < r_{pbis}$ by alpha 2% the test items as valid. The formula used is as follows:

$$rpbis = \frac{\overline{Xb} - \overline{Xs}}{SDt}\sqrt{\frac{p}{q}}$$

Description :

$\overline{Xb}$ = The average score of students who answered correctly

$\overline{Xs}$ = The average of students who answered incorrectly

SDt = Standard deviation

p = Proportion of correct answers to all the students' answers

q = 1 – p

Reliability Test

Reliability is the level or degree of consistency of an instrument. According Arikunto (2009: 173) The main purpose of calculating the reliability is knowing the level of precision and objectiveness test scores. Reliability index ranges between 0 - 1. More higher the coefficient of reliable test, so the precision is more higher too. The reliable test is the test that has steadily score, relative unchanged even test given in different situation and different time. Instead, the tests are not reliable like rubber to measure the length, measurement results can be fickle rubber (inconsistent) To determine the coefficient of reliability tests of multiple choice questions, the formula used Kuder Richardson 20 (KR-20), namely:

$$KR-20 = \frac{k}{k-1}\left[1 - \frac{\sum p(1-p)}{SD^2x}\right]$$

Keterangan :

k = Number of items

SD^2x = variance

Method

This experiment is descriptive research. Descriptive research is research that aims to describe a independent variable even only one variable or more [5]. So the researchers did not make a comparison of these variables in other samples, or look for relationships between variables. The study was conducted to determine the quality of a multiple-choice test items in logical mathematics material at X grade of Pangeran Antasari Medan senior school academic years 2015/2016. The population conducts of 30 students of X grade of Pangeran Antasari Medan senior school academic years 2015/2016. The variables in this study are reliability test, distinguish power, difficulty levels, validity, and reliability multiple choice test item academic years 2015/2016.

Discussion

Of the 40 items of daily test multiple choice made by teachers of mathematics courses, do an analysis to see the validity of the test. The results of the analysis of the validity of the test are presented in Table 1.

Tabel.1. Validity test Item Analyzing

Test item	Validity	r_{tabel}	Interpretation	Test item	Validity	r_{tabel}	Interpretation
1	0,440096	0,361007	V	21	0,396227	0,361007	V
2	0,555857	0,361007	V	22	0,445331	0,361007	V
3	0,297494	0,361007	TV	23	0,555857	0,361007	V
4	0,111137	0,361007	TV	24	0,139804	0,361007	TV
5	0,446483	0,361007	V	25	0,49084	0,361007	V
6	0,272229	0,361007	TV	26	0,480782	0,361007	V
7	0,168346	0,361007	TV	27	0,546915	0,361007	V
8	0,291405	0,361007	TV	28	0,564651	0,361007	V
9	0,620997	0,361007	V	29	0	0,361007	TV
10	0,490794	0,361007	V	30	0,589212	0,361007	V
11	0,263161	0,361007	TV	31	0,311584	0,361007	TV
12	0,490718	0,361007	V	32	0,343517	0,361007	TV
13	0,47038	0,361007	V	33	0,440672	0,361007	V
14	0,134635	0,361007	TV	34	0,529796	0,361007	V
15	0,41703	0,361007	V	35	0,526575	0,361007	V
16	0,598579	0,361007	V	36	-0,04429	0,361007	TV
17	0,455526	0,361007	V	37	0,309617	0,361007	TV
18	0,05897	0,361007	TV	38	0,294421	0,361007	TV
19	0,334916	0,361007	TV	39	0,2242	0,361007	TV
20	0,620997	0,361007	V	40	0,295299	0,361007	TV

Based on the data above it can be conclude that, from 40 multiple choice item, 20 item are valid, and 18 item are invalid. Valid test items can presented in table 2.

Table 2. Valid Test Item

NO	Item test	Validity	r_{table}	Interpretation	NO	Item test	Validity	r_{table}	Interpretation
1	1	0,4401	0,36101	V	12	21	0,39623	0,36101	V
2	2	0,55586	0,36101	V	13	22	0,44533	0,36101	V
3	5	0,44648	0,36101	V	14	23	0,55586	0,36101	V
4	9	0,621	0,36101	V	15	25	0,49084	0,36101	V
5	10	0,49079	0,36101	V	16	26	0,48078	0,36101	V
6	12	0,49072	0,36101	V	17	27	0,54692	0,36101	V
7	13	0,47038	0,36101	V	18	28	0,56465	0,36101	V
8	15	0,41703	0,36101	V	19	30	0,58921	0,36101	V
9	16	0,59858	0,36101	V	20	33	0,44067	0,36101	V
10	17	0,45553	0,36101	V	21	34	0,5298	0,36101	V
11	20	0,621	0,36101	V	22	35	0,621	0,36101	V

The result of reliability test are presented in table 3

Table 3. Analysis of Reliability Test

K	22
Var t	20,80575
Total pq	3,29
KR-20	0,88196
INTERPRETATION	RELIABILITY TEST SO HIGH

By using the formula KR-20, as well as calculations using excel help with questions that otherwise valid number as many as 22 items, the obtained reliability coefficient of 0.88196. Based on the criteria of reliability coefficient can be concluded reliability of the test used is very high. This means that the test has a high consistency, the test does not change even tested on the different situation and the different time.

The result of Difficulty Index Analyzing test items are presented in table 4

Table 4. Difficulty Index Analyzing

ITEM NUMBERS	IK	INTERPRETATION	ITEM NUMBERS	IK	INTERPRETATION
1	0,2	HARD	21	0,7	EASY

2	0,86667	EASY	22	0,93333	EASY
5	0,63333	MEDIUM	23	0,86667	EASY
9	0,86667	EASY	25	0,9	EASY
10	0,56667	MEDIUM	26	0,76667	EASY
12	0,86667	EASY	27	0,6	MEDIUM
13	0,83333	EASY	28	0,9	MUDAH
15	0,9	EASY	30	0,83333	MUDAH
16	0,76667	EASY	33	0,83333	MUDAH
17	0,83333	EASY	34	0,83333	MUDAH
20	0,86667	EASY	35	0,4	MEDIUM

From the analysis we concluded that 17 test items have an index of difficulty easy categories, four test items have difficulty index of the medium category, and one of the test items have difficulty index difficult category. The results of the analysis of distinguish power items are presented in tables 5

Table 5. Distinguish Power Analyzing

Ordinal Number	Items Number	Distinguish Power	Interpretation	Ordinal Number	Items Number	Distinguish Power	Interpretation
1	1	0,4	Enough	12	21	0,66667	Good
2	2	0,733333	Good	13	22	0,73333	Good
3	5	0,6	Good	14	23	0,73333	Good
4	9	0,733333	Good	15	25	0,73333	Good
5	10	0,533333	Good	16	26	0,66667	Good
6	12	0,733333	Good	17	27	0,6	Good
7	13	0,733333	Good	18	28	0,73333	Good
8	15	0,733333	Good	19	30	0,73333	Good
9	16	0,733333	Good	20	33	0,73333	Good
10	17	0,733333	Good	21	34	0,73333	Good
11	20	0,733333	Good	22	35	0,46667	Good

From the analysis we concluded that 21 of the test items having distinguishing good category, and one of the test items having distinguishing good enough category.

Results distractor analysis of the options used are presented in Table 6

Table 6. Distractor Analiyzing

	ITEMS NUMBER																					
Key Answer	1	2	5	9	10	12	13	15	16	17	20	21	22	23	25	26	27	28	30	33	34	35
	E	B	C	D	E	D	A	C	B	D	C	C	A	C	A	A	D	D	B	D	A	E
A	2	2	0	1	2	1	25	1	2	0	1	3	28	2	27	23	0	1	0	3	25	13
B	22	26	0	0	1	1	3	1	23	1	2	5	0	1	1	1	6	0	25	0	1	2
C	0	2	19	1	3	2	1	27	2	4	26	21	1	26	0	0	5	0	2	0	1	0
D	0	0	10	26	7	26	1	0	3	25	1	0	0	1	2	5	18	27	2	25	3	3
E	6	0	1	2	17	0	0	1	0	0	0	1	1	0	0	1	1	2	1	2	0	12

While the percentage of the distractor used option is presented in Table 7

Table 7. Distractor Percentage

	ITEMS NUMBER																					
%	1	2	5	9	10	12	13	15	16	17	20	21	22	23	25	26	27	28	30	33	34	35
	E	B	C	D	E	D	A	C	B	D	C	C	A	C	A	A	D	D	B	D	A	E
A	6,7	6,7	0,0	3,3	6,7	3,3	83,3	3,3	6,7	0,0	3,3	10,0	93,3	6,7	90,0	76,7	0,0	3,3	0,0	10,0	83,3	43,3
B	73,3	86,7	0,0	0,0	3,3	3,3	10,0	3,3	76,7	3,3	6,7	16,7	0,0	3,3	3,3	3,3	20,0	0,0	83,3	0,0	3,3	6,7
C	0,0	6,7	63,3	3,3	10,0	6,7	3,3	90,0	6,7	13,3	86,7	70,0	3,3	86,7	0,0	0,0	16,7	0,0	6,7	0,0	3,3	0,0
D	0,0	0,0	33,3	86,7	23,3	86,7	3,3	0,0	10,0	83,3	3,3	0,0	0,0	3,3	6,7	16,7	60,0	90,0	6,7	83,3	10,0	10,0
E	20,0	0,0	3,3	6,7	56,7	0,0	0,0	3,3	0,0	0,0	0,0	3,3	3,3	0,0	0,0	3,3	3,3	6,7	3,3	6,7	0,0	40,0

While the interpretation of the option used distractor is presented in Table 8

Table 8. Distractor Interpretation

Interpretation of Distractor Result	ITEMS NUMBER																					
Options	1	2	5	9	10	12	13	15	16	17	20	21	22	23	25	26	27	28	30	33	34	35
A	Y	Y	N	N	Y	N	**0**	N	Y	N	N	Y	**0**	Y	**0**	**0**	N	N	N	Y	**0**	Y
B	Y	**0**	N	N	N	N	Y	N	**0**	N	Y	Y	N	N	Y	N	Y	N	**0**	N	N	Y
C	N	Y	0	N	Y	Y	N	**0**	Y	Y	**0**	**0**	N	**0**	N	N	Y	N	Y	N	N	N
D	N	N	Y	**0**	Y	**0**	N	N	Y	**0**	N	N	N	N	Y	Y	**0**	**0**	Y	**0**	Y	Y
E	**0**	N	N	Y	0	N	N	N	N	N	N	N	N	N	N	N	N	Y	N	Y	N	**0**

Description :
Y = Distractor Functionally
N = Distractor Dis-functional
O = Key answer

From the table above it can be concluded that of the 22 valid test item, there is a distractor which is functioning properly.

Conclusion

Based on the data above it can be conclude that, from 40 multiple choice item, 20 item are valid, and 18 item are invalid. The coefficient of reliability test is 0,88196. Based on reliability criteria the coefficient of reliability test is very high. That means the test has a high consistency, the test does not change even the situation and distinguish time. 17 item tests have an index of difficulty easy categories, and 4 item tests have an index of difficulty medium categories, and 1 item test have an index of difficulty difficult categories. 21 test items having distinguishing good category, and 1 of the test items having distinguishing good enough category. From 22 item valid test, the distractor has a good function. Then it can be concluded from the test item 40 item 22 item created test items can be stated already have a good quality test.

REFERENCES

1. Aiken, L. R. (1987). *Assessment of Intelectual functioning.* Massachussetts: Allyn and Bacon Inc.
2. Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, California: Brookd/Cole Publishing Company.
3. Anas Sudijono. (2005). *Pengantar evaluasi pendidikan.* Jakarta: Raja Grafindo Persada.
4. And T. Kellaghan (Eds.), *Evaluation Models* (pp. 16–32). Boston, MA: Kluwer Academic Publishers.
5. Arifin, Zainal. (2009). *Evaluasi Pembelajaran*. Bandung: Rosda.
6. Arikunto, Suharsimi. (2009). *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara
7. Asmawi Zainul dan Noehi Nasoetion. (1997). *Penilaian Hasil Belajar*. Pusat Antar Universitas, Direktorat Jenderal Pendidikan Tinggi: Departemen Pendidikan Dan kebudayaan.
8. Cronbach, L., (2000), “Course Improvement through Evaluation”, dalam D. I. Stufflebeam, G. F. Madoux.
9. Feldt, L. S. & Brennan, R. L. (1989). “Reliability” dalam Linn R. L. (Eds.), *Educational Measurement Third Edition.* (pp. 105-146). New York: McMillan.
10. Fernandes, H. J. X. (1984). *Testing and measurement.* Jakarta: National Education Planning, Evaluation and Development.
11. Kumaidi. (1994). *Studi analitik terhadap karakteristik internal dari ujian seleksi masuk ke perguruan tinggi.* Makalah disajikan dalam seminar pengkajian ujian saringan masuk ke perguruan tinggi di BALITBANG Depdiknas Jakarta.
12. Messick, S. (1989). “Validity” dalam Linn, R. L. (Eds.), *Educational measurement third edition.* (pp. 13-103). New York: McMillan.
13. Rasyid, H., dan Mansur, (2008), “Penilaian Hasil Belajar”, Bandung: Wacana Prima.
14. Saifuddin Azwar. (2003). *Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar.* Yogyakarta: Pustaka Pelajar.
15. Sax, G. (1980). *Principles of educational and psychological measurement and evaluation.* Belmont: Wadsworth Publishing Company.
16. Sudjana, Nana. (2009). *Penilaian Hasil Proses Belajar Mengajar*. Penerbit Remaja Rosdakarya. Bandung.
17. Sumadi Suryabrata. (2004). *Pengembangan alat ukur psikologi.* Yogyakarta: Penerbit Andi.
18. Sumiati dan Asra, (2008), “Metode Pembelajaran”, Bandung: Wacana Prima
19. Yen, W. M. (1992). “Item Response Theory”. dalam Alkin M. C. (Eds.), *Encyclopedia of Educational Research* (pp. 657-666). New York: Macmillan Library Reference USA.