AN ANALYSIS OF STUDENTS READING FINAL EXAMINATION BY USING ITEM ANALYSIS PROGRAM ON ELEVENTH GRADE OF SMA NEGERI 8 MEDAN

Dewi manalu

English Department,University of HKBPNommensen,
Sutomo St. No.4A, Medan,Indonesia
Email:dewimanalu54@gmail.com

ABSTRACT

The purpose of this study was to determine the quality of the reading final examination in SMA N8 Medan grade eleventh in terms of reliability, level difficulty, discrimination power and level of distractor. This research is qualitative-quantitative research. The subject of research are the grade XI of SMA N8 Medan. Data is analyzed by ANATES program version 4.0.9. The analysis shows that: (1) items of multiple choice question that can be said as valid are 14 items (56%), while the invalid items amounted to 11 items (44%). (2) Items mutiple choice question can said as reliable because it equal 0,90 so it can said reliable. (3) items of multiple choice questions that categorized easy are 3 items (12%), satisfactory category 7 (28%) difficult category 2 (8%) and the other calculate categorized very easy 3 items (12%) and very difficult 3 (12%). (4) items of multiple choice questions that categorized poor are 12 items (48%),average category are 2 items (8%), good items category 1 items (4%) and excellent items are 8 items (32%).

Keywords: Test Item Analysis, Reliability, Level of Difficulty, Discrimination power, Levelof Distractor.

Introduction

Education is needed by everyone, it can be said that education is experienced by all people. Education plays an important role in ensuring the survival of the nation and the state. Education as a means to improve the Human Resources (HR). Human Resources quality can bring the progress of Science and Technology in all aspects of life and bring people to a more advanced civilization and logical mindset (Rukmana, 2013). Human resource development can be enhanced by improving the quality of education in schools, which is determined by various factors. These factors are teachers, students, environment, infrastructure, learning time and learning process.Human Resources quality will appear along with a good quality of education as well. Until recently, the education component that is used, needs to be improved Especially in the evaluation system of learning outcomes. Evaluation of education according to UU No. 20 Year 2005 are the activities of controlling, underwriting and, determining the educational quality of the various components of education at every track, level and type of education as a form of educational responsibility. One of the components that have to be evaluated is the learning outcomes of students. According to Sridadi (2007) assessment of student's learning outcomes intended as an attempt to collect a variety of continuousand comprehensive information about the process and the outcomes of learning that have been achieved by the students through the teaching and learning activities as the basis for determining the next steps. Based on Government Regulation (PP) No. 19, 2005 on National Education Standards, particularly Article 63 paragraph 1 states that education in primary and secondary level consists of assessment of learning outcomes by educators, educational units and government. Article 64 Paragraph 1 states that the assessment of learning outcomes as intended in Article 63 paragraph 1, point (a) is done on an ongoing basis to monitor the process, progress, and improvement of the results in the form of daily tests, midterm examination, the examination in the end of the semester, and examination in order to go to the next grade.

Evaluation is a process of collecting data to determine the extent in terms of what and how the educational goals have been achieved (Suharsimi, 2013: 3). Evaluation can be used as a benchmark in decisions making about the object that will be evaluated. Measurement is an attempt to compare a particular casewith so its quantitative (Arikunto and Jabar, 2004). While the assessment is making a decision on something in terms of good or poor, healthy or sick, clever or stupid, high or low, and so

on (Djaali and PudjiMulyono, 2008). One technique that is commonly used for the assessment is by using the test.

The test is a procedure that can be used to determine or measure something in accordance to the way and the rules that have been setSuharsimi, 2013: 46), while the non-test is a procedure used to measure the affective domain such as attitudes, interests, talents , and motivation, for example using questionnaires, interviews, observations, and others (Sudijono, 2011: 67). Although there are two kinds of measuring instruments evaluation activities, but the test is often used for the evaluation tool. The test in question is a test that is made by the teacher *(teacher-made test)*. The results of the test should reflect the real situation, because the results of the test will be used to make decisions. The size of the mistake (error) could affect the measurement of learning outcomes assessment, the test would be said as a good questions if they meet the requirements like eligibility, level of difficulty, distinguishing power, the pattern of answer distribution and relationship or correlation of each item with an overall score. Besides, the test should also have the characteristic of validity, reliability and objectivity. To find those things, it is needed a valuation activity through the test item analysis to obtain information about the test which has been meets the requirements of a good question. Tests are said to be good as a measurement tool if they meet the requirements of the test, which has: validity, reliability, practicability, objectivity and economical (Suharsimi, 2013: 72).

According to Suharsimi (2013: 73), the data can be said to be valid if it is in accordance with the actual situation. If the data is valid, the instrument used is valid, because it is able to provide a description of the data correctly according to the actual situation. Test is said to be reliable if the test has been given repeatedly and provide consistent results (Suharsimi, 2013: 74). If thereis no subjectivity factor in the scoring system, the test can be said as objective. Tests were carried out should be practical and easy for its administration. The test is easy to implement, easy to correct, and guided by clear instructions that can be given or initiated by others. The test is said to be economical if the implementation does not require high cost, a lot of labour, and a long time (Suharsimi, 2013: 77). Tests that already meets some of the requirements that have been set, will be used as a measurement tool of student's achievement and learning success. Through the test item analysis, it will be obtained information about the good and not good items. Good items will be retained and stored in the question bank, while the not good items should not be used again in the next assessment. The activity of revised the test items were performed in order to make the test qualified enough to be use ase a measurement tools of student's learning outcomes. The test item analysis is done by calculating the aspect of validity, reliability, level of difficulty, distinguishing, and the effectiveness of detractors. (ZainalArifin, 2011: 22

Language is a set of rule. Realization of rule can be recognized through grammar. Sipayung(2018) stated that Teaching English is a complex enterprise. He is doing it in classroom. In a simple synoptic, a teacher is greeting students, explaining the matter, ask learners participants and do the homework. This has been teaching trade and a tradition. But, when a search of teaching is done, it is an interction between a teacher and his community, the classroom subject, the pupil. Teacher made test as a level of students understood. One aspect in learning process is do an evaluation to the students. It is done to make sure whether the laerning process has been running well over the term. In education field, evaluation has an important role, because it shows the results of learning program. The objective of the evaluation itself is to help the teacher ascertain the degree to which educational objectives have been achieved, to review the effectiveness of teaching method and to help the teacher know his pupils as individuals.

According to Ahman and Glock evaluation is the systematic process of determining the effectiveness of educational endeavors in the light of evidence.. once assesment information is collect, teacher use it to make a decisison, refelection or judgment about pupils, instruction or classroom climate.The evaluation is defined as the process in making decision and solution for education process based on the result of tests, other assessments, and others reports Brown & Priyanvada(2010:9). For English language education, the evaluation is conducted in many aspects of education such as curriculum, strategies of teaching, references and test items. The good evaluation is believed will have good benefits in national education because by evaluating the system of education, the government can improve the quality of education system.

Developing a test is a complex and reiterative process which subject to revision even if the items were developed by skilful item writers. Many commercial test publishers need to conduct test analysis, rather than trusting the item writers judgement and skills to improve the quality of items that need to be proven statitically after trying out was performed. This tudy is a part of test

development process which aim to analyse the reading test items.

Testing refers to an effort to measure the result of students learning in teaching learning process. Consequently, the teachers should have an ability to arrange and good test and analyze of a test. Therefore, the accuracy and the carefulness of teachers may have a big impact on the increase the quality of teaching particularly in giving the judgement of students ability. This information is ver useful for both students in their learning and the teacher in their learning. It can be feedback for the teacher, who have responsibility to meet the instructional objectives, while for getting the data for the sake of evaluation purpose, one them is by using a test. The test items are supposed to be well constructed so it canbe used efficiently. To be an effective test, it has to fulfill the critieria of a good test, they are validity, rehability and practicality. The test is recognized as valid if can measures what supposed to be measure, it illustrates their performance.Related to the importance of the evaluation, it is necessary to consider that the test should be well constructed. As a means of evaluation, a test is administrated to get information about the students improvement and to measure the result of the teaching learning process. Test a test activity which is held at theend of teaching learning process in one semester. That is why, that test is a kind of test which is intended as a feedback from thestudents and also as a result of teaching from the teachers in one semester. Thisinformation will be used to consider and to decideseveral rules not only for thestudent's but also for the teachers in increasing the quality of teaching learningprocess. And the English test is made by MGMP (*Musyawarah Guru Mata Pelajaran*). While MGMP itself consists of a team who hasresponsibility to design a test for each subject, it means that the semester testitems are rarely analyzed by theteachers afterthey are tested.

To analyze the semester test items, there are some criteria of a good test accordingto some expert. A good test should have (1) Validity, (2) Reliability, (3) Level ofdifficulty, (4) Discrimination Power, and (5) The Quality of Options. Thisresearch was concerned with the whole with test items designed by MGMP. Thisincludes test analysis and item analysis. Test analysis is administered to determineand describe such criteria as face validity, content validity, construct validity, andreliability. And the item analysis is used to determine about the level of difficulty,discrimination power, and the quality of options. Shohamy (1985:3) supports that a test is a sample of knowledge and needs to be agood representation of it. It means that, what should be tested just a sample ofbehavior or knowledge, not the whole or behavior what the teachers has taughtand the students have learned because it is also impossible to measure all of thestudents' abilities. The things that should be taken into account is the sample mustbe representative in the sense which is tested, it should reflect the knowledge thathas been taught. The test that has been analyzed was achievement test and it wasdesigned byMGMP. Achievement test tried to investigate the students'achievement based on the objective of a given material. Achievement test(Harrison as quoted by Hayatunnisa, 2003:8) tries to evaluate the test takers'language in relation to a given curriculum or material which the test-taker hadgone through in a given course. It is intended to show the standard which thestudents have reached in relation to other students at the same stage.

A good test should fulfill certain the criteria. There are four criteria of a good testaccording to some expert; they are validity, reliability, level of difficulty, anddiscrimination power. Validity refers to the extent to which an instrument really measures the objectiveto be measured and suitable with the criteria based on Hatch and Farhady(1982:250). Inother words, a test can be said to be valid to the extent that it measures what it issupposed to measure. If the test is not valid for the purpose for which its design,the scores do not mean what they are supposed to mean. Reliability refers to theconsistency of measurement that is, to see how consistent test scores or otherevaluation results are from one measurement to another (Gronlund, 2000:193). Itmeans that a test is administered to the same condition on different occasion, theextent that it produces different result, it is not reliable. Discrimination power isan aspect of item analysis, discrimination power tells about which is the itemdiscriminates between the upper group students and the lower group students.Shohamy (1985:81) states that discrimination index tells about the extent to whichthe item differentiates between high and low students on that test.

Difficulty level is one of kind of item analysis. Level of difficulty was concernedwith how difficulty or easy the item for the students. Shohamy (1985:79) statesthat difficulty level relates to how easy or difficult the item is from the point ofview of the students who took the test. It is important since test items which aretoo easy can tell usnothing about differences within the test

population. If theitem too easy, it means that most or all of the students obtained the correctanswer. In contrast, if the item is difficult, it means that most or all of the studentsget it wrong. The quality of options is a distribution of test in decidedalternatives on a multiple choice test. It is obtained by calculating the number oftest who choose the alternatives A, B, C, or D or those who do not choose anyalternatives. From this way, the teachers would be able to identify whetherdistracters function well or bad.

A good evaluation (test) is important instrument to know whether the students need a help or not. Test have been widely use to demonstrate level of proficiency of the students, and at the same time function as policy instruments to implement educational standards stated by Phakiti & Roever (2011 : 29). This study is a part of test development process which aims to analyse the reading comprehension test items. A teacher who is competent in terms of pedagogics must be able to make quality questions. The quality of the questions is viewed from the aspects of difficulty items, item discrimination power and level of distractors.

Item analysis is a crucial part in a test development process as it functions to provide information about items that should be improved in terms of quality for laters tests or even be eliminated due to misleading.According to Gronlund (1977) there are some benefit to do analysis of the test items. First, it is provide useful information for class discussion of the test, second, it provides data that helps the students improve their learning. Third, it provides insight and skills that lead to the preparation of better testin the future. Item analysis is process which examines students responses to individual test items (questions) in order to asses the quality of those items and of the test as a whole. It will show whether the test appropriate or not for the students. Item analysis is especially valuable in improving item which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. In addition, item analysis is valuable for increasing instructors skill in instruction , and identifying spesific areas of course which need greater emphasis clarity.

Reading comprehension can be defined as the ability to understand vocabulary in order to paraphrase and make a summary of infromation from the text based on Manarin et al (2015). It is the activity to reconstruct a message from written symbols to a form of language, and it involves inference. Some consideration are needed in many cognitive processes and combines both decoding process and inferential activity and so that the readers can really comprehend the text. In the context of English as a foreign language, reading English textbooks becomes a big issues, particularly for students with non-English background. Reading skills seems to be the big problem to the students because most of them find reading English text is difficult . reading comprehension is part of critical reading that can be a determinant to academic success. Lowes et al (2004) argues that reading is essential to understand basic concept of a subject to gather information for completing assignments, and to improve English skill, particularly to increase vocabulary. One of many characteristics of reading at higher education is critical reading which involves such features as identifying patterns of textual elements, distinguishing main and supporting ideas, making credible evaluation and arguments and making relevant inference about the text.

According to Brown test is a tool that serves to make decisions related curriculum and other areas. Brown (2004) points out that a test is a way of measuring one's ability, knowledge or performance in a given domain. In language testing context, most test measure test takers competence, such ability to perform language skill to speak, write, listen or read to one subset of language. These performance based test sample the test takers actual use of language which infers general competence. A test of reading comprehension, for example, may consist of several short reading passages each followed by a limited number of comprehension questions. From the result of the test the examiner may infer a certain level of general reading ability. Constructing a test is not a simple task. It invvolves a science and art of many complex task, such a planning, test preparation an administration, scoring, statistical analysis, and test result report (brown, 2004; downing, 2010). One of cruscial stages in test development is statistical analysis of a test.

Assessment is a systematic approach to collecting information and makin inferences about the ability of a student or the quality or success of a teaching course on the basis of various source of evidences. To collect the information of the students improvement in reading comprehension, some assessment of reading are essential to be carried out. Unlike speaking and writing, the reading process and product cannot be seen and observed spesifically. For this reason, all assessment of reading must be carried out by designing assessment of reading, such as the types

of genres of written text, the components of reading ability and spesific tasks. Furthermore, the types of reading performance will influence the assessment tasks as well. Brown lists a number of possible tasks for assessing perceptive, selective, interactive and extensive reading. Related to the spesifications of reading in this study that is assessing students interactive reading performance so there are three types of possible assessment tasks applied as follow 1. Multiple-Choice Multiple choice in this study provides not only the vocabulary and grammatical items but also the context to assess the students understanding of information in the text. The context is presented by putting a pair or part of a text followed by questionsin which the studentshave to respond correctly.2. Impromptu Reading Plus Comprehension This type of assessment involves impromptu reading and responding to questions. It is commonly used in proficiency test. In this test, students are provided a reading passage followed by questions and have to responds to the items. The set of questions in impromptu reading covers the comprehension of some features of reading : (1) main idea (2) expressions/idioms/phrases in context (3) interference (4) grammatical vocabulary in context. These spesifications and the questions are in line with strategies of effective reading :skimming for main idea, scanning for details, guessing word from context, inferencing using discourse makers etc which are assessed in this study. 3. Short Answer Task In this type of assessment, reading passage is presented, and the students read questions that must be answeredin a sentences or two. The question might cover the same spesifications similar to the impromptu reading. Those three types of assessments task have the combination of form-focused and meaning-focused objective. They cover the objective of reading assessment especially in the comprehension issues and embody the evidences of students reading comprehension.

Method

This study aimed particularly process of analysing multiple choice questions in reading skills, to improve item quality. Twenty-five of multiple choice question had been tested to get the evidence on item quality. The items were one correct answer type, having a stem and four options, one of them being correct and the other three being "distractor". Based on the students responses, the test items were then analysed using ANATES Software. Anates is a software for statistic calculation. Anates V4 develop by Drs.Kanoto M.Pd and the programer is Yudi Wibisono ST. This software very effective to calculate about reliability, item difficulty, item discrimination power, and level of distractor. The advantages in using this program is this program can be use to analyze the test item of multiple choice and it really helps for investigate the data.

Analysis data refers to find and set sistematically the data which have collected from the collecting data process to be easily to understand in this study. The analysis quantitatively is doing to order investigate the test items from students answer or response and the key answers. Anates is a software for statistic calculation. Anates V4 develop by Drs.Kanoto M.Pd and the programer is Yudi Wibisono ST. This software very effective to calculate about reliability, item difficulty, item discrimination power, and level of distractor. The advantages in using this program is this program can be use to analyze the test item of multiple choice and it really helps for investigate the data.

How to use anates software : 1) Open the application and the apllication will ask how many question and how many students then fill it. 2) Input the data / score students one by one then save. 3) Click on " process data" on the option. 4) Then data will process and give the result based on item difficulty, item discrimination and level of distractor that you choose before.

Findings and Discussion

Based on the analysis, the reliability was 0,90, the value was categorezed as a high it has a certain criteria in test reliability. In testing item difficulty, the formula were applied :

$$P = \frac{B}{JS}$$

Based on analysis results show that the number of multiple choice categorized as easy questions amounted to 3 items (12%), satisfactory category 7 (28%) difficult category 2 (8%) and the other calculate categorized very easy 3 items (12%) and very difficult 3 (12%).

Table 1.
Distribution of multiple choice question of final examination of SMA Negeri 8 Medan based on level difficulty

No	Difficulty level	Item number	Total	Percentage
1	0,70 - 1,00 (easy)	4,16,21	3	12%
2	0,31 – 0, 71 (satisfactory)	1,2,5,8,12,17,20,24,25	9	36%

3	0,00 – 0, 30 (difficult)	9,11		2	8%
4	(-) – 0,00 (very difficult)	13,14,15		3	12%
5	1,00 – up (very easy)	3,6,7,10,18,19,22,23		8	32%

Item Discrimination

Item discrimination has a significant role to examine if an item is low or high quality. In testing item discrimination, the formula were applied :

$$D = \frac{BA}{JA} - \frac{BB}{JB} = PA - PB$$

Where :

D = Discrimination Power

JB = the number of students in lower group who tried the item

PB = index of difficulty in lower group

BA = the number of students in upper group who answer correctly

BB = the number of students in lower group who answer correctly

PA = index of difficulty in upper group

Based on the analysis result, the number of multiple choice questions which have poor categories are 12 items (48%),average category are 2 items (8%), good items category are 1 items (4%) and excellent items are 8 items (32%).

Table 2.
Distribution of multiple choice question of final examination of SMA Negeri 8 Medan based on discrimination power

No	Difficulty Level	Item Number	Total	Percentage
1	0.00 – 0.19 (Poor)	3,5,6,7,10,13,14,15,16,18,19,23	12	48%
2	0.20 – 0.39 (average)	4,22	2	8%
3	0.40 – 0.69 (good)	9	1	4%
4	0.70 – 1.00 (excellent)	1,2,8,11,12,17,20,25	8	32%

Distractor Analysis

The distractors can function well if at least chosen by 5% of all learners who participate in test. Criteria for assessing the use of distractors can adapt based on Arikunto (2005 :220) are :

a. Acceptable, because it has been chosen at least 5% the total number of students who try to answer. A good distractor can distract more students from lower group than students from upper group.

b. Refused, because it does not exist by the students to choose the distractor.

c. Revision, because it attracts more students from upper group than students in lower group. It is revision distractor if choose by less of 5% of the total number of students who try to answer. A distractor can be called revision only on the sentence structure. It can be re-write the effective revision.

For example :

Once, there was a family who had a baby. They also had a dog named Blackie and it was very smart and faithful. Blackie used to take care the baby while the family working in the Farm.

One afternoon, while they were working, they heard Blackie barking. Blackie was running toward them with mouth covered in blood. The husband shocked. He thought that Blackie killed the bab. Suddenly he took his grass knife and hit the dog until it died.They got home quickly and saw the baby was sleeping. When he looke around, he found a die big Snake covered in blood. It seemed that blackie killed the Snake and he killed blackie.

1. Who was Blackie? Blackie was…

   a. The dog   c. The wife

   b.The husbandd.   d. The baby

Option	Upper	Lower
A*	8	19
B	0	1
C	0	0
D	0	0
Total	8	20

Distractor B are functioning poorly since it attracts only one students from lower group . These distractor are able to distract more students in lower group. Distractor C and D is unless attracts because no students in both of this group choose the option. This the analysis of distractor:

a. Distractor B = revise

b. Distractor C = refused

c. Distractor D = refused

Conclusion

Item analysis has provided useful information about the characteristic of items in one test. Some items, after the analysis, might be revised, changed or even removed. Based on the analysis above, it is found that the test had high reliability with coefficient 0.90. based on the finding, many items were categorized as marginal and poor category in terms of the level of difficulty, discrimination power and level of distract.

References

Arifin, Zaenal. (2012). *Evaluasi Pembelajaran.* Bandung : PT. Remaja Rosdakarya.

Arikunto, S. Jabar, C. (2004). *Evaluasi Program pendidikan.* Jakarta : Bumi ksara.

Brown. (2000). *Principles of Language Learning and Teaching.* San Francisco : Longman.

Brown. (2001). *Teaching by Principle An Interactive Approach to Language Pedagogy.* San Francisco : Longman.

Brown. (2004). *Principles and Classroom Practices.* New York : Pearson Education.

Brown & Priyanvada. (2010). *Language Assesment :Principles and Classroom Practices.* New York : Pearson Education.

Danuwijaya. (2018). *Item Analysis of Reading Comprehension Test for Post-Graduate Students.* English Review : Journal of English Education.

Day & Bamford. (1998). *Extensive Reading in the Second Language Classroom.* Cambridge : Cambridge University Press.

Downing, S. M. (2010). Test Development. *In International Encyclopedia of Education (3rd ed. P.159-165).* Elsevier

Djaali and M. Pudji. (2008). *Psikologi Pendidkan.* Jakarta : Pustaka Pelajar

Gronlund, N.E. 1977. *Constructing Achievement Test, 2nd Edition.* New York : Prentice Hsll, Inc. University of Illinois.

Hatch, E, and H. 1982.*Research Design and Statistic for Applied Linguistic*.London: New Burry House, inc.

Hayatunnisa, 2003.*An Analysis of the First Semester English Test for the SecondYear Students of SMU AL-KAUTSAR Bandar Lampung*.UnpulishedScript. Bandar Lampung: FKIP University of Lampung

Henning, G. 2000. *A Guide to Language Testing*.Mass : New Bury HousePublishers.

Heaton. (1990). *Classroom Testing (Longman Keys to Language Teaching) The Fast Free Shipping.* Florida : Pearson Education Limited.

Hughes. (2003). *Testing for Language Teacher.* Cambridge : Cambridge University Press.

J. Stanley Ahman and Marvin D.Glock.(1967). *Evaluating Pupil Growth.* Boston : Allyn and Bacon, Inc p.4

Jhon. (2009). *Merancang Kuesioner.* Depok : Universitas Gunadarma.

Lind & Grondlund. (2009). *Measurement and Assesment in Teaching.* University of Florida : Pearson.

Lehr. F & Osborn, J. (2001). *Put Reading First :The Research Building*

Manarin. K, Carey, Rathburn, & Ryland. (2015). *Critical Reading in Higher Education: Academic Goals and social Engangement.* Bloomington, Indiana: Indiana University Press

Nunan, David. (1989). *Designing Tasks for the Communicative Classroom.* Cambridge University Press : Library of Congress Cataloguing in Publication Data.

Nurgiyantoro. (2001). *Penilaian Dalam Pengajaran Bahasa dan sastra.* Yogyakarta : BPFE.

Phakiti & Roever. (2011). *Currents Issues and Trends In Language Assesment in Australian and New Zealand.* New Zealand : Australian University.

Purwanto. (2011). *Evaluasi Hasil Belajar.* Yogyakarta : Pustaka Pelajar.

Richard & Schmidt. (2002). *Dictionary of Language teaching and Applied Linguistics.* England : Longman.

Rukmana, Nur Hidayati Indra. 2013. *Analisis Butir Soal Ujian Akhir Semester Gasal Mata Pelajaran Teori Kejuruan Kelas XI Akutansi SMK YPKK 1 Gamping.* Skripsi. Yogyakarta : FE UNY

Sari, Y. A. (2016). Item Analysis of English Mid-Term Test Items for the Second Semester of the Seventh Grade Students of SMP Negeri 2 Wonosari in the 2015/2016 Academic Year. Surakarta : Islamic Institute of Surakarta.

Sipayung, K.T. (2018). Generic Teaching Skills in English Language Teaching. *International Journal on Studies in English Language and Literature (IJSELL),* 6(7), p1-15.

Suharsimi, Arikunto. (2013). *Dasar – Dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara.

Shohamy, E. 1985.*A Practical Handbook inLanguage Testing For the SecondLanguage Teacher*. Tel Aviv: Tel Aviv University.

Sridadi, Bambang. (2007). Diktat Mata Kuliah Evaluasi Pembelajaran Penjas. Yogyakarta : FIK UNY

Thompson & Levitov. (1985). *Assesing Students Performance Using Test Item Analysis and Its Relevance to the State Exit Final Exams.* Kendall Campus : Miami Dade College.