ABSTRACT

RAFIKASARI, RIZQI SAKTI. 2015, An Analysis Of Multiple Choice Test Items Used In English Try Out At Ninth Grade Of Smp N 2 Jetis Ponorogo In Academic Year 2014/2015, Thesis, English Education Department, Tarbiyah Faculty, State Islam College of Ponorogo, Advisor Pryla Rochmawati, M.Pd.

Key Word: Try Out Test, Multiple Choice.

Test is one of the instruments to measure the ability of the students. By the test we can know who are the clever students and the less bright students, so we can know what we must do to the students. One of kind of the test is try out test.Try out test is important as a footstep for student to pass real test on UAN. Most of try out test is in the form of multiple choice .Therefore, it considered important to evaluate the quality of test**.** Reffering to the explanation above, this reserach was intended to analyze the validity, reliability, item difficulty and item descrimination in test item used in try out test. How does the multiple choice test meet the requirement of good test in term of validity ?How does the multiple choice test meet the requirement of good test in term of reliability ?How does the multiple choice test meet the requirement of good test in term of item difficulty ?How does the multiple choice test meet the requirement of good test in term of item descrimination ?

This design of the research was descriptive quantitative research. The data of this research was answer sheet from 92 students who participated the try out test in 2014.This research takes place at SMP N 2 Jetis. The researcher used random sampling, were taken 72 amswer sheets as sample. The technique of data collection used documentation of English answer sheet in try out test from the teacher. The data were have analyzed using quantitative descriptive. The testing research used product moment formula for validity, ^{split} half formula for reliability,FV $\frac{r}{n}$ for item difficulty, and D= $\frac{BA}{JA} - \frac{BB}{JB}$ = PA –PB for item discrimination fomula.

The result of the reserach showed that Multipe choice test has low validity. It was drawn from the result of analysis 22 items classified as valid criteria, and 28 items classified as invalid criteria. Reliability of multiple choice test classified reliability. The scale Alpha showed reliability of the test with 0,86. From the result showed that reliability. This test has medium difficulty items, the easy item showed 34% too easy item, 66% medium item, and 0% difficult item. In which, there are 31 items from 50 items have medium criteria. This test has fairly bad discrimination items. It is drawn from the results of analysis that showes 52% items is classified as fairly bad items. Where there are 26 from 50 items that includes fairly bad criteria. Then, 48% is classified as bad item.It means that, try out of SMP N 2 Jetis has good test.

CHAPTER I

INTORDUCTION

A. Background of the study

Evaluation can be defined as the systematic gathering of information for the purpose of making decisions[1]. Without information gathered during the process of teaching, it is unable to say way learners have done better. Information is needed by teacher, education, curriculum, etc. The purpose of decision making is done to get comparison and selection the result of evaluation. The decision of evaluation is made to make the suitable of evaluationis instrument.

Evaluation is defined here as the systematic attempt to gather information in order to make judgements or decision.[2]

Evaluation will help to show students how to identify the learning strategies they used for a recently completed learning task, encourage studentsto reflect on their own learning processes, plain activities in which students evaluate the effectiveness of the learning strategies they have used for a specific task, asses how effectly students are appliying the strategies taught, include learning strategies evaluation assessment portofolios, evaluate your own learning strategies instruction.[3]

[1] Lyle F. Bachman, Fundamental Cosiderations in Language Testing (New York : Oxford University Press, 2003), 22.

[2] Brian K. Lynch, Language Program Evaluation:Theory and Practice, (Cambridge University Press, 1996). Page 2

[3] Anna Uhl Chamot, The Learning Strategies Handbook (Longman, 1999) page 114

Good test provides the opportunity for learnerto show how much they know about language structure and vocabulary, as well as how they are able to use these formal linguistic feature to convey meanings in classroom language activities through listening, speaking, reading and writing. Commonly, test is divided into two categories, they are objective and subjective test. Objective test usually has only one correct answer, (or, at least, a limited number of correct answer), they can be scored mechanically.[4]

Subjective test in literature means essay examination. This test demands the students to explain, discussion and compare with the reasons, in this section demands the ability of students in expressing of opinion in writing language. In this research only uses objective test. There are some kinds of objectives test, such as multiple choice, true-fals, completion and matching. It is useful at this stage to consider multiple choice items in some details as they are undoubtedly one of the most widely that uses types of items in objecyive test. "Soal pilihan ganda adalah bentuk tes yang mempunyai satu jawaban *yang benar atau yang paling benar"*.[5] Each multiple choice item should have only one answer. This answer must be absolutely correct unless they instruction specifies choosing the best option (as in some vocabularies tests). Although this may seem in easy matter, it is sometimes extremely to construct an item having only one correct answer.

Test is one of the instruments to measure the ability of the students. By the test we can know who are the clever students and the less bright students,

[4] J.B. Heaton, Writing English Language Test (London and New York: Longman), 25

[5] Nana Sudjana, Penelitian Hasil Proses Belajar Mengajar (Bandung: Rosdakarya, 2009), 48.

so we can know what we must do to the students. There are many kinds of test. According to Arthur Hughes considering the purposes for which language testing is carried out. It goes on to make a number of distictions : between direct and indirect testing, between discrete point and integrative testing, between norm-referenced and criterion-referenced testing, and between objective and subjective testing. Finally there is a note on communicative language testing.

The tests can be categorized into good test if it confirm several consideration, namely validity, reliability, and practically.

Briefly, te validity of a test is the extent to which it measures what it is supposed to measure and nothing else. Every test, whether it be a short informal classroom test or a public examination, should be as valid as the constructor can make it. The test must aim to provide a true measure of the particular skill which it is intended to measure to the extent that it measures external knowledge and other skills at the same time, it will not be a valid test. Validity is unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree which that evidence supports the inferences that are made from the score. The inference regarding specific uses of a test are validated, not the test itself.[6]

The next requirement to be a good test is reliability. The investigation of reliability is concerned with answer and question, "how much of an individual's test performance is due measurement error or to factor other

[6] Lyle F Bachman, Fundamental Concideration In Language Tasting. (New York : Oxford University)237

than the language ability we want to measure ?" And with minimizing the effect of these factors on tests score.[7]

In practically test has many categories, namely item difficulty, item discrimination, item distracter. But in this research just take two categories of tests, they are item difficulty and item discrimination. Item diffficulty has to do with how easy (or difficult) an item is from the viewpoint of the group of the student.[8].

Difficulty level is to find the level of difficulty for each question. This is simply the precentage of students (high and low combined) who got each question right. Generally, a test question is considered too easy if more than 90 percent get it right. An item is considered too difficult if fewer than 30 percent get it right. (You can see why noting that a person might get 25 percent on a four-option test by guessing). Referring to the example, we find that item 1 is acceptable.[9]

The type of tests based scoring method, that they are objective test, matching test, true-false test, multiple choice test, and essay test. But in this research analyze the test using multiple choice. Multiple choice in some detail as they are undoubtedly one of the most widely types of items in objective test. Multiple choice item is one of the most difficult and time consuming types of item construct, numerous poor multiple-choice test

[7] Lyle F Bachman, Fundamental Concideration In Language Tasting. (New York: Oxford University)161

[8] John W. Oller, Language Test at School Pragmatic Approach (Mexico : University of New Mexico),246

[9] Harold S. Madsen, Techniques in testing, (Oxford University Press . New York , 1983)page 181-183.

now abound. The chief critism of the multiple choice item, however is that frequently it does not lend itself to the testing of language as communications. The optimum number of alternative, or options, for each multiple-choice item is five in most public tests.

Reffering to the result of observation in SMP N 2 Jetis at 10th April 2015, the tests that ussualy at school in final examination is try out test.

Try out test is important as a footstep for student to pass real test on UAN. It considered important to know the quality of the test. Therefore , the writer is interested in conducting the research during with the quality of the try out test in terms of its validity, reliability, item difficulty, and item discrimination. From explanation above , the writer considered the need for research on " An Analysis of Multiple Choice Test Used In English Try Out Test At Ninth Grade of SMP N 2 Jetis Ponorogo In Academic Year 2014/2015.

B. Focus of Research

To avoid a far-ranging discussion, this study focuses on an analysis validity, reliability, item difficulty, item discrimanation on english try out test for ninth grade of SMP N 2 Jetis in academic year 2014/2015.

C. Statement of the Probem

1. How does the multiple choice test meet the requirement of good test in term of validity ?
2. How does the multiple choice test meet the requirement of good test in term of reliability ?
3. How does the multiple choice test meet the requirement of good test in term of item difficulty ?
4. How does the multiple choice test meet the requirement of good test in term of item descrimination ?

D. Objectives of the Study

Based on the problem statements, this reasearch had some obejectives described as follow :

To analyzed the validity, reliability, item difficulty and item descrimination in test item used in try out test.

E. Significance of the Study

The reasearcher hopes the result of this study is a valuable one. It is expected to be any theoritically and practically. It is described as follow :

1. Theoritical significance
   a. The results of this study could knowed the validity,reliability, item difficulty, item descrimination of english try out test.
2. Practical significances

a. The result of this study are expected to be useful for the study of the teacher to know the validity, reliablity, item difficulty and item descrimination of the test items in english try out test

b. The results of this study will be useful to reader, especially teachers of English lesson and understand the valid of the test items.

F. Organization of the Thesis

The thesis is divided into five chapters that can be presented as follows:

Chapter I gives introduction that contains the background of the study. Focus of research. The statement of the problem, the objectives of the study, the significance of the study and organization of the thesis.

Chapter II presents review of related literature. In this chapter writer tells about theoritical background, priview study, and theoritical framework.

Chapter III is reasearch methodology. In this chapter writer tells about researc design, population and sample, instrument of data collection, technique of data collection and technique of data analysis.

Chapter IV is researcher result. In this chapter the writer tells about reasearch location, data description and analysis, and discussion.

Chapter V gives the conclussion and suggestions.

CHAPTER II

REVIEW OF RELATED LITERATURE

A. Evaluation

1. Definition of Evaluation

Evaluation may be defined as a systematic process of determining the extent to which instructional objectives are achieved by pupils. There are two important aspects of this definition. First, note that evaluation implies a systematic process, which omits casual, uncontrolled observation of pupils. Second, evaluation assumes that instructional objectives have been previously identified. Without previously determined objectives, it is difficult to judge clearly the nature and extent of pupil learning[10]. Evaluation is defined here as the systematic attempt to gather information in order to make judgements or decision.[11]

It is important at this point to note that the evaluation is not something which only takes place summatively, at the end of a course of instruction. Informal monitoring should, in fact, be happening right trhough the course. Any element in the curriculum process may be evaluated, as any may affect learner progress, and it is up to individual teachers and curriculum personnel to decide how widely they should cast the net. Some obvious candidates for evaluation are initial planning

[10] David Nunan, The Learner-Centred Curriculum, (New York, Cambridge University Press, 1990). 119.

[11] Brian K. Lynch, Language Program Evaluation:Theory and Practice, (Cambridge University Press, 1996). 2

procedures, program goals and objectives, the selection and grading of content, materials and learning activies, teacher performance and the assessment process itself as well as learner achievement.[12]

Evaluation will help to show students how to identify the learning strategies they used for a recently completed learning task, encourage studentsto reflect on their own learning processes, plain activities in which students evaluate the effectiveness of the learning strategies they have used for a specific task, asses how effectly students are appliying the strategies taught, include learning strategies evaluation assessment portofolios, evaluate your own learning strategies instruction.[13]

B. Test

1. Definition of Test

Testing is a universal feature of social life. Throughout history peole have been put to the test to prove their capabilities or to establish their credintials; this is the stuff of Homeric epic, of Arthurian legend, in modern societies such tests have proliferated rapidly [14]. A test plain words, is a method of measuring a person's ability or knowledge in a given domain. The definition captures the essential components of test. A test is first a method it is a set techniques, and item that constitute an instrument of some sort that requires performance or activity on the part of the test taker are. Also being certain competence in a test is ability or competence.

[12] David Nunan, The Learner-Centred Curriculum, (New York, Cambridge University Press, 1990). 119.
[13] Anna Uhl Chamot, The Learning Strategies Handbook (Longman, 1999) . 114
[14] Tim McNamara, Language Testing. (Oxford University Press. 2000). 3

A test sample performance but infers certain competence [15]. Testing is about making inferences ; this essential point is obscured by the fact that some testing procedures, articularly in performance assessment, appear to involve direct observation.[16]

Teachers often regard testing as a matter for 'experts' outside the class, and not something that they can be involved in themselves. This unit sets out to show the importance to teachers of being aware of different testing techniques, and shows how tests can be used informally in the class to give useful information to both the teacher and the students. Testing techniques are often similar to teaching techniques, but with a different purpose.[17]

2. Types of Tests

a. The Types of Tests Based Scoring Method

This scoring can be done according to one of two fundamentally different ways based on the level of objectivity of scoring, the types are in the following :

1) Objective test

Objective test is a test that can be done by scoring his high level of objectivity. The resulting scores at the end of scoring against a participant test work objective is basically no different and will be the same if the scoring is done by two or more

[15] Brown H Douglas, Principle of Language Learning and Teaching, 384
[16] Tim McNamara, Language Testing. (Oxford University Press. 2000) 7
[17] Adrian Doff, Teach English A training *course for teachers Trainer's* Handbook (Cambridge University Press, 1988) 257

corrector, or by the same collector who does scoring twice or more at different times,

2) Test match

Test match gave the task to the participants of tests to match or match (matching) two-part test in terms of content or the meaning of the two parts logically interrelated.

3) Ttrue-false test

True False test consists of a number of test items, each form pernyataan.beberapa between the statement is true in a sense as it should be, some other form of false statements, which do not conform or contrary to it should be.

4) Multiple-choice test

Multiple choice test is a kind of objective test that each of its test items have more than two options. different from the true-false test that his test items generally consists of only one statement to be tagged as right or wrong, one item or multiple-choice test consisting of a basic statement or question, followed by several statements that match or selection of the correct answer.

5) The subjective test

Tests categorized as a subjective test scoring when test participants work can not be done objectively and can only be done subjectively. questions and tasks administered in the test was formulated in such a way as to invite the response and execution of

tasks that test participants varied in focus, content, wording, and the length of the short answer. Such answers can only be balanced in accordance with the opinions and subjective assessment of a proofreader. in case of a job-takers subjketif in check by two or more different corrector, the results of his assessment is very likely to be different between one corrector with other corrector.

6) Test essay

In general aged essay test is used to refer to the subjective test in general. which as previously disclosed, his scoring can only be done subjectively. more specifically, the test essay refers to tests that answer the essay or process of writing description in different styles, such as set of problems descriptive and argumentative, in accordance with the subject.

7) Test questions using question words

Subjective tests of this type consist of grains test formulated in the form of interrogative sentence which begins with a question word. in English, such interrogative sentence known as wh-questions because the sentence structure that uses and begins with the words written by wh (wh-questions words), such as who, what, why, where, when, which.

8) Test short answer questions

As well as the test with questions beginning with question words, the kind of subjective test short answer questions test

consists of test items, each in the form of questions that are formulated using question words, generally one of wh-questions words.

9) Test complements

Complete test consists of test items, each shaped like a short discourse sentence, which must be equipped by the test participants in the empty part of the original text, whether in the center, at the beginning, or the end of a sentence.

The Types of Test Based on The Preparation Development Based on the way the preparation and the way of development, the test can distinguish between, (1) the test terstandart (standarized test) and (2) teacher-made tests (teacher-made test).

a. Standarizedtest

The preparation and development of standarized test with the characteristics of good was done through various stages with attention and fulfill a number of requirements. Stages of preparation and development that includes the preparation of a blueprint (blueprint), the preparation and writing of test items, some of the pilot phase, several stages of revision, and the study of the characteristics of a good test, analysis of test items, and especially the validity and reliability. In the blueprint outlined the steps that will be in use which consists of:

a) Planning tests

b) Writing tests

c) Critical Review

d) Trial

e) Revision

f) The final form

b. Teacher-made tests

Compared with tests terstandart development and preparation of teacher-made tests much simpler. Similarly, implementation procedures. Everything is done by teachers themselves are assumed to have sufficient knowledge on the need to develop a good test. As someone who constantly meet and interact with the learners for a certain period of time, teachers not only know their habits dn tingah behavior, but also the level of capability in addition to their advantages and disadvantages.

c. The Type of Test based on The Score Interpretation

Tests can be distinguished from one another based on references used in interpreting the resulting scores.

d. Norm Reference test

On the use of norm reference test, a participant test scores are interpreted by comparing the scores obtained all the other participants that have the same spelling test. Scores were

interpreted as a reflection of the level of mastery of the knowledge test participants or ability being measured by the test in question. In a norm reference test, the level of mastery of the participants in the comparison tests among participants test based on a norm internally calculated on the basis of the acquisition of scores of all participants on the same tests.

e. Criterion Reference test

On the use of reference criteria for the interpretation of the test scores are generated based upon a criterion, namely the minimum skill level predetermined as an indicator of the mastery of the target field tests.[18]

b. The Types of Tests Based on the Function

In learning known for a number of types of tests relating to the existence and use phases and functions in its implementation. In relation to the stage of completion of planned learning can be distinguished their formative tests, summative tests, pretest and posttest.

1. Formative tests

Formative test target is the level and quality of the learning achievement of the participants of the learning objectives have been occur phase of the implementation of a specific formative tests. In addition to the level of achievement of the participants, the

[18] M. Soenardi Djiwandono,Tes Bahasa : Pegangan bagi pengajar bahasa, (Jakarta:2008) 35-78

results of formative tests also provide information about which parts of the learning materials to a certain extent that has been conveyed and possessed well by learning and other parts that have not reached the expected level of mastery.

2. Summative tests

Summative tests conducted before or at the end of the implementation of the learning program is part of a comprehensive evaluation of the success of the whole learning program implemented. As a means of a thorough evaluation of the overall success of the program the target summative tests include a mastery level of learning of all material that has been planned and implemented over a certain period as a quarterly, semester, or a year, and others.

Pretest

In connection with the implementation of a learning program, pretest held before or at the beginning of the implementation of a program of learning.

Postes

As mentioned previously, postes held before or at the end of the implementation of the learning program.[19]

[19] Ibid, 91 - 94

c. Based on the Purpose

Reffering purpose, although in other cases the purpose may affect the form. The most familiar distiction in term of test purpose is that between achievement and proficiency test[20]. According to Arthur Hughes, Type of test based on the purpose are in the following:

1) Proficiency tests

Proficiency tests are designed to measure people's ability in a language regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of language courses which people taking the test may have followed. The function of this these tests is to show whether candidates have reached a certain standart with respect to certain specified abilities. Such examining bodies are independent of the teaching institutions and so can be relied on by potential employers etc. To make fair comparisons between candidates from different institutions and different countries. Despite differences between them of content and level of difficulty, all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken.

2) Achievement test

Most teachers are unlikely to be responsible for proficiency tests. It is much more probable that they will involved in the

[20] Tim McNamara, Language Testing. (Oxford University Press. 2000) 6

preparation and use of achievement tests. In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how sucessfull individual students, groups of students, or the courses themselves have been in achieving objectives. In the view of some testers, the content of a final achievement test should be based directly on a detailed course syllabus-content approach. It has an obvious appeal, since the test only contains what it is thought that the students have actually encouratered, and thus can be considered, in this respect at least, a fair test. It has been argued in this section that it is better to base the content of achievement tests on course objectives rather than on the detailed content of a course. However, it may not be at all easy to convince colleagues of this, especially if the latter approach is already being followed. Not only is there likely to be natural resistance to change, but such a change may represent a threat to many people.

3) Diagnostic tests

Diagnostic tests are used to identify students strengths and weaknesses. They are intended primarily to ascertain what further teaching is necessary. At the level of broad language skills this is reasonably straightforward. We can be fairly confident of our ability to create tests that will tell us that a student is particularly weak in, say, speaking as opposed to reading in language. The laackof good

diagnostic tests is unfortunate. They could be extremly useful for individualised instruction or self-instruction. Learners would be shown where gaps exist in their command of the language, and could be directed to sources of infromation, excemplification and practice. Happily, the ready availability of relatively inexpensive computers with very large memories may change the situation. Well-written computers progammes would ensure that the learner spent no more time without the need for a test administrator. Tests of this kind will still need a tremendous amount of work to produce. Whether or not they become generally available will depend on the willingness of individuals to write them and of publisher to distribute them.

4) Placement tests

Placement tests, as their name suggests, are intended to provide information which will help to place students at the stage (or in the part) of the teaching programme most appropiate to their abilities. Typically they are used to assign students to classes at different levels. The placement test which are most succesful are those constructed for parcticular situations. They depend on the identification of the key taylor-made rather than bought off the peg. This usually means that they have been produced in house. The

work that goes into their construction is rewarded by the saving in time and effort trhough accurate placement [21]

3. Criteria of Good Test

The good test confirm validity, reliability, and practically.

a. Validity

Briefly, the validity of a test is the extent to which is measures what it is supposed to measure and nothing else. Every test, whether it be a short, informal classroom test or a public examination, should be as valid as the constructor can make it [22]. A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. With wich it is meant to be concerned. It is obvious that a grammar test, for instance, must be made up of items testing knowledge or control of grammar. But, this in itself does not ensure content validity.

1. Content Validity

What is the importance of content validity? First, the greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure. A test in which major areas identified in the specification are under-represented or not represented at all is unlikely to be accurate. Secondly, such a test is likely to have a harmful back wash effect. Areas which are not tested are likely to become areas ignored in

[21] Ibid , 9-14
[22] J. B Heaton, Writing English Language Test, (London and New York 1988) 159

teaching and learning. Too often the content of tests is determined by what is easy to test rather than what is important to test. [23].

2. Criterion-related Validity

Criterion related validity refers to the realationship between the score on a measuring instrument and an independent external variable (criterion believed to measure directly the behavior or characteristic in question. Some writer make distinction between two types of criterion-related validity is predictive validity and concurrent validity. Both are concerned with empirical relationship between test score and criterion, but distinction is made on the basic of the time when criterion data are collected. Concurrent validity is concerned wth the correlation between test score and a criterion measure available at the same ar a very close in time. Predictive validity is concerned with the correlation between test scores and a criterion that occurs at a latter point in time.

3. Constract Validity

Constuct validity is concerned with the extent to which a test measure a specific trait of construct. It is the type of validity that is essential for tests that are used to asses individuals on certain psycological traits and abilities. The term construct is used

[23] Arthur Hughes, Testing for Language Teachers, (Cambridge University Press.1989).22.

refers to something that is not itself directly measureble but which explain observable affefcts.[24]

A teacher needs to be satisfied that a particular test is an adequate definition of a construct. Lets say you have been given a procedure for conducting an oral interview. The scoring analysis for the interview weighs several factors into a final score: pronounciation, fluency, grammatical accuracy, vocabulary use, and sociolinguistic appropiateness. The justification for these five factors lies in a theoritical construct that claims those factors as major components of oral proficiency. So, on the other hand, if you were asked to conduct an oral proficiency interview that accounted only for pronounciation and grammar, you could be justifiably suspicious about the construct validity of such a test. Validity is a complex concept, yet it is indispensable to the teacher's under standing of what makes a "good" test. If in your language teaching you can attend to the practically, reliability, and validity of test of language, whether those tests are classroom tests related to a part of a lesson, or final exams, or proficiency tests, then you are well on the way to making accurate judgment about the competence of the learners with whom you are working.[25]

[24] Winston Rinehart and Holt, Introduction to Research in Education, 197-201

[25] Brown H Douglas, Teaching by Principles an interactive approach to language pedagogy (San Fransisco,) 389.

Table 2.1
Kinds of Validity

Type	Meaning	Procedure
Content validity	How well the sample of tasks represents the domain of tasks to be measured	Compare the test tasks to the test specifications describing the task domain under consideration,
Criterion related validity	How well test performance predicts future performance or estimate current performance on some valued measures other that the test self.	Compare test score with another measure of performance obtained at a later date (for prediciton) or with another measure of performance obtained concurrently (for estimating present status)
Construct validity	How test performance can be describe psycollogical	Experimentally determine what factors influence scores on the test.[26]

b. Reliability

Reliability is a necessary characteristic of any good tests : for it to be valid : all a test must first be reliable as a measuring instrument. If the test is administered to the same candidates on different occasions (with no language practice work taking place between these occasion), then, to the extent that is produces differing result, it is not reliable. Reliability is of primary importance in the use of both public achievement and proficiency tests classroom tests [27]. But if this is the case, it would seem to imply that we can never have complete trust in any set of test scores. We know that the scores would have been different if the test had been administered on the previous or the following day. This is inevitable, and we must accept it. What we have

[26] David Nunan, the Learner-Centred Curriculum , 120.
[27] J. B Heaton, Writing English Language Test, (London and New York:Longman) 162.

to do is construct, administer and score tests in such a way that the scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but a different time. The more similar the scores would have been, the more reliable the test is said to be.

$$r_n = \frac{N}{N-1}\left(1 - \frac{m\ (N-m)}{Nx^2}\right)$$

where N = the number of items in the test :

m = the mean score on the test for all the testees

x = the standart deviation of all the testees' scores

r_n = reliability.[28]

c. Practically

1) Item Difficulty

The index of difficulty (or facility value) of an item simply shows how easy of difficult the particular item proved in the test. The index of difficulty (FV) is generally expressed as the fraction (or precentage)of the students who answered the item correctly. It is calculated by using the formula :

[28] Ibid 164

$$FV = \frac{R}{N}$$

R represents the number of correct answers and N the number of students taking the rest. Thus, if 21 out of 26 students tested obtained the correct answer for one of the items, that item would have an index of difficulty (or a facility value) of 77 or 77 per cent.[29]

2) Discrimination Index

Item discrimination (ID) is a statistic that indicates the degree to which an item separates the students who performed well from those who did poorly on the testas a whole. These two groups are sometime referred to as the "high" and "lower" scorers or "upper" and "lower" proficiency students. The reason for identifying these two groups is that ID allows teachers to constrast the performances of the lower students. The process begins by determining which students had scores in the top group on the whole test and which had scores in the bottom group. The upper and lower groups are sometimes defined as the upper and lower third or 33 percent. Some test developers will use the upper and lower 27 percent.[30]

The discrimination index of an item indicates the extent to which the item discriminates between the testees, separating

[29] J. B Heaton, Writing English Language Test, (London and New York: Longman), 175-176

[30] James Dean Brown, Testing in language Program,(New York, 2005) 68.

the more able testees from the less able. The index of discrimination (D) tellss us whether those students who performed well on the whole test tended to do well or badly on each item in the test. It is pressupposed that the total score on the test is a valid measure of the student's ability (i.e the good students tends to do well on the test as a whole and the poor student badly). On this basic, the score on the whole test is accepted as the criterion measure, and it thus becomes possible to separate the 'good' students from the 'bad' ones in performances on individual items.

$$D = \frac{Correct\ U - Correct\ L}{n}$$

(D = Discrimination Index ; n = Number of candidates in one group

U = Upper half and L = Lower half. The index D is thus the different between the proportion passing the item in U and L).[31]

There is one caution in applying discrimination to our language tests. When doing an item analysis of rather easy and rather difficult question, be careful not to judge the items too hasrhly. For example, when almost 90 percent get an item right, this means that nearly all low students as well as high students have marked the same (correct) option. As a result,

[31] J. B Heaton, Writing English Language Tests, (London and New York)180

there is little opportunity for a difference to show up between the high and low groups. In other words, discrimination is automatically low. Also be careful when evaluating very small classes- for example, those with only 10 or 12 students. This is especially true is students have been grouped according to ability. Occasionally even useless items like this can be revised and made acceptable. For example, an evaluation of one overseas test found a question with unacceptable discrimination.[32]

4. Try out test

a. Definition of try out test

Trying out or field-testing items can provide extremely useful information during the tests development process. When conducting an item tryout, use a sample of examinees similar to those who will take an assessment once it is administered operationally (for official score-reporting purposes). This step is particularly important for items that will be used with ELLs.

b. Purposes of Item Tryouts

There may be several reasons to conduct item tryouts. Data may be collected in order to:

1) inform decisions about how appropriate the items are for a sample of examinees similar to the operational population,

[32] Harold S. Madsen, Techniques in testing, (Oxford University Press . New York , 1983)181-183.

2) inform content and fairness reviews of the items,
3) evaluate timing requirements for new or existing item types,
4) evaluate the clarity of instructions to examinees,
5) support the scaling or equating of test forms,
6) inform the standard setting process by providing performance data, which panelists will receive as feedback on cutscores, on different groups, and
7) assess whether ELLs of different proficiency levels can understand the text of the items.This is important when English language proficiency is not the construct of interest.

c. Types of Item Tryouts

Item tryouts may take several different forms, ranging from one-on-one interviews with students, through small-scale pilot tests, to large-scale field tests. As with other activities described within these guidelines, it may not be possible to implement each of these types of item tryouts in a given testing program because of resource constraints. However, we describe them here so that readers can make informed decisions about when and whether each type may be useful.

1) One-on-One Interviews

One-on-one interviews with students who have been administered the items can provide much useful information. These interviews can take the form of informal debriefings after students have completed the tasks, or more formal cognitive laboratory activities

where students are interviewed either while they are answering the questions or afterward

2) Small-Scale Pilot Tests

Small-scale pilot tests may also provide useful information on how students respond to the items. In this data collection format, test developers administer the items to a larger sample of students than is used for one-on-one interviews, and, generally, one-on-one debriefing does not take place

3) Large-Scale Field Tests

In large-scale field tests, test developers administer the items to a large,representative sample of students. Because of the size and nature of the sample, statistics based on these responses are generally accurate indicators of how students may perform on the items in an operational administration. If the tryout items are administered separately from the scored items, motivation may affect the accuracy of the results. When the tryout items are embedded among the scored items, students do not know which items count and which do not, so motivation is not a factor. [33]

5. Multiple Choice

Multiple choice in some detail as they are undoubtedly one of the most widely types of items in objective test. Multiple choice item is one of the most difficult and time consuming types of item construct, numerous

[33] Educational Testing Service(ETS), Guidelines for the Assessment of English Language Learners,2009 .23-24

poor multiple-choice test now abound. The chief critism of the multiple choice item, however is that frequently it does not lend itself to the testing of language as communications. The optimum number of alternative, or options, for each multiple-choice item is five in most public tests.

6. Previous reserach

In this study discusses about analysis validity and reliability of English try out test, the researcher looks at the previous of study from:

Binti Lailatul Munawaroh with the title : "An analysis of multiple choices item test on English of MTs Ma'arif Munggung Pulung. The researcher analyzed item difficulty, item discrimination and item distracter and reliability using program ITEMAN version 3.0 because program ITEMAN easy to use and simple. Data included this program are number of question, answer key, answer choice, student's name and student answersw of test result.

The result of this research, The writer drwas the conclusion that the test have low validity. It was drawn from the results of analysis item difficulty showed 48% item classified as bad item. And 52% item did not show item distratcter. The scale Alpha showed reliability of the test with 1,99. From the result showed that, the test has low validity doesn't mean that the test also has low reliability. This research used research evaluation method.

And other research was taken from Fina Syuriah Rahayu with the title : "An Analysis of Validity and Reliability on English Midterm of

SMP 2 Kauman Ponorogo. The researcher analysis of item difficulty, item discrimination and item distracter and reliability using program ITEMAN version 3.0 because program ITEMAN easy to use and simple. Data included this program are number of question, answer key, answer choice, student's name and student answer of test result. The result of this research is this test have low validity, analysis item difficult showed 82,5% item classified as easy item. Item discrimination showed 30% item classified as very bad item. And 77,5% item showed item disctracter not work. Reliability of this test classified has high reliability is the scale Alpha is 0,678.

And other research was taken from Fitra Khoirul Anwar with the title : " The Effectiveness of English Item Test in Final-Test on Odd Semester (Item difficulty and Item Discrimination Level) at XI IPS English Student of SMA Bakti Ponorogo. The researcher analysis of item difficulty and Item Discrimination level. The designe of this research was descriptive research. This test has good term of difficulty items. The result of analysis that showed 67,5% was classified as good item. In which there were 27 items from from 40 items had good criteria. Then, 27% was classified as very difiicult item and 5% was classified as very easy items. So, this test is good test which whic this test is not too easy and is not too difficult. This test had satisfactory item discrimination level. It was drawn from the result of analysis that showed 32,5% satisfactory criteria. Then 22,5% was classified as good item, 22,5% was classified as poor item,

15% was classified as bad item, and 7,5% was classified as excellent items.

The different of the research with the research above is the research describe the formula validity, reliability, item difficulty, and item discrimination, the research demonstrated how to calculate tha value of validity, reliability, item difficulty, and item discrimination, to calculate the formula of validity and reliability using computer program, and for item difficulty and item discrimination manually without using computer program.

7. Theoritical Framework

Evaluation of teaching learning process can be administrated in the form of good test. A good test must be consider the validity and reliability. So, the test has been qualified good test, it is necessary to the review of the item tests.

A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. With wich it is meant to be concerned. It is obvious that a grammar test, for instance, must be made up of items testing knowledge or control of grammar

for it to be valid : all a test must first be reliable as a measuring instrument. If the test is administered to the same candidates on different occasions (with no language practice work taking place between these occasion), then, to the extent that is produces differing result, it is not reliable.

In practically of the test in this research has item difficulty and item discrimination. The index of difficulty (or facility value) of an item simply shows how easy of difficult the particular item proved in the test. For the item discrimination these two groups are sometime referred to as the “high” and “lower” scorers or “upper” and “lower” proficiency students.

CHAPTER III

RESEARCH METHOD

A. Reserach Design

Desain penelitian merupakan sebuah rencana prosedural yang menjadi panduan peneliti untuk menjawab pertanyaan-pertanyaan peneliti secara valid, obyektif, akurat dan ekonomis.[34] This statement can be explained that research design is the procedurals palnning which it is research guidance validly objectively, accurately, and economic.

The data in this research were numeric form. The researcher categories this research as the quantitative research. Quantitative research in this study include in descriptive research. There is definition about descriptive research," penelitian deskriptif adalah suatu metode penelitian yang ditunjukkan untuk menggambarkan fenomena-fenomena yang ada, yang berlangsung pada saat ini atau saat yang lampau.[35] It can be said that descriptive research is a research method that addressed to describes the exist phenomena, and it occurs in present or in past time. Other definition said that the research limited on trying to express the problem, condition, and event naturally. So it just expresses the fact finding.[36].

[34] Restu Kartiko Widi, Asas Metodologi Penelitian(Yogyakarta: Graha' ilmu, 2010), 212.

[35] Nana Syaodih Sukmadinata, Metode Penelitian Pendidikan (Bandung: PT. Remaja Rosdakarya,2009), 54.

[36] Hadari Nawawi, Metode Penelitian Bidang Sosial (Yogyakarta: Gadjah Mada Univercity Press, 2007), 33-34

B. Research Setting

This research takes place at SMP N 2 Jetis. It is located at Ngasinan village. The selection the school in SMP N 2 Jetis is schools get accredits B. It enable the research to analysis validity, reliability, item discrimination, and item difficulty of try out test.

C. Population and Sample

1. Population

Population is defined as all members of any well defined class of people, event or subject.[37] According to Sugiyono *"populasi adalah* wilayah generalisasi yang terdiri atas: objek/subjek penelitian yang mempunyai kualitas dan karateristik tertentu yang ditetapkan oleh *peneliti untuk dipelajari dan kemudian ditarik kesimpulanya."* [38]

From the definition above, the researcher summarizes that population is all of the subjects who will be researched. The population in this research is all the ninth grade students at SMPN 2 Jetis Ponorogo in academic 2014/2015.

All of the ninth grade students in SMPN 2 Jetis in academic 2014/2015 is 92 students, it is can be the population of the research.

[37] Fred N. Kerlinger, Foundations of Behavioral Research (New York: Holt Rinehart and Winston,1996),52

[38] Sugiyono, Metode Penelitian Kuantitatif Kualitatif dan R&D (Bandung:Alfabeta,2008),80.

2. Sample

Sample is small group that is observed or sample is a portion of a population. Sample is collection of elements or individuals that are part of the population. The sample is smaller than the total of population.

Therefore, sample must have characteristic to be possessed by the population. Whether a sample is a good representation for the population is highly dependent on the extent to which the characterictic of the sample was equal to the population characteristics. Because the research analysis is based on sample data while the conclusion will be applied to the population it is important to obtain samples for represented population.[39]

In this research, the researcher determined the subjects as a sample are the ninth grade students of SMPN 2 Jetis in academic year 2014/2015, which consists 4 classes they are 9A.B.C.D, every class consists of 30 students. The technique of this sampling is Random Sampling.So, the sample in this research is 72 students.

D. Technique of data collection

In this reserach, the researcher used documentation in technique of data collection. "Teknik dokumentasi adalah cara mengumpulkan data melalui peninggalan tertulis, seperti arsip-arsip dan termasuk juga buku-

[39] Sugiyono, Metode Penelitian Pendidikan (Bandung: Alfabeta,2008),188.

buku tentang pendapat, teori, dalil, atau hukum-hukum dan lain-*lain"*.[40] It can be said that documentation technique is the ways of collecting data through writing thing, like files, theories, concepts, laws and etc that relate with research.

Is basically a way to find the data information by reading note, journal, and everything that can make to be the data. In this study, documentation is used to look for the English test paper in English Try out Test in academic year 2013/2014. The data includes of the question, answer from all students, and the key answer of the test in ninth grade students. The total of objective questions is 50 items number and the testees of the student are 98 respondents.The steps of collecting data in this study are :

1. The researcher collected the question sheet and answer sheets form student of Engliosh Try Out Test, VI, VII, and VIII students of SMP N 2 Jetis ponorogo in academic Year 2013/2014.
2. Processing data with study of answer sheet of student. Where the number items are true or false in each student and it is corrected with the answer key.
3. The researcher analyzed the data.

[40] S Margono, Metodologi Penelitian Pendidikan. (Jakarta: Rineka Cipta, 2003) 181.

E. Technique of Data Analysis

According Arikunto, data analysis technique is a process to analyze the data found. To prove is the hypothesis which formula by the researching is right or wrong.[41]

This research is done trhough the preparation and implementation steps. The preparation which determine the place of research. The reserach took in SMP N 2 Jetis. The implementation step was done by taking data to be studied, then the preformed an anlysis of data obtained to be made about the quality description of items, and will be presented in the from of discussion of research result.

Analysis of item test was conducted quantitative analysis. Quantitative analysis performed with validity, reliability, item difficulty, item discrimination beacuse the researcher wish know about multiple choice item at English try out test at SMP N 2 Jetis

1. Validity and Reliability

a. Validity

Validity means the extent to which inferences made from assesment results are appropriate, meaningful, and useful in terms of the purpose of the assessment.[42] Test is said to have validity if the result are in accordance with the criterion, in term of parallels between the results of tests with the criterion. Technique used are

[41] Arikunto, Procedure Penelitian Suatu Pendekatan Praktik, 156

[42] H. Douglas Brown, Language Assessment Principle and Classroom Practice, (New York: Longman,2000),22.

usually the product moment correlation technique proposed by Pearson. The formula is:

Rxy : $\frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2]\,[N\sum Y^2 - (\sum Y)^2]}}$

Rxy = product moment correlation

∑X = sub of all values of X

∑Y = sub of all values of Y

∑XY = amount of product between the X and Y values

N = Number of students

b. Reliability

A reliable test is consistent and dependable.[43] Reliabilty is the degree to which a test consistently measures whatever it measures. Reliabilty indicates the extent to which individual differences on test scores are attributable to true differences versus chance errors.

In this research used Spearman Brown (Split half) formula to measure the reliability of the test. The formula is [44] :

$r_i = \frac{2rb}{1 + rb}$

r_i = the internal reliability of all the instrument

r_b = the correlation of product moment between the first half and the second half

[43] H. Douglas Brown, Language Assessment Principle and Classroom Practice, (New York: Longman,2000)hal 20.

[44] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif dan R&D,(Bandung:Alfabeta,2006),131.

2. Item Difficulty and Item Discrimination

a. Item Difficulty

The index of difficulty (or facility value) of an item simply shows how easy of difficult the particular item proved in the test. The index of difficulty (FV) is generally expressed as the fraction (or precentage)of the students who answered the item correctly. It is calculated by using the formula :

$$FV = \frac{R}{N}$$

(FV) = is generally expressed as the fraction (or precentage)of the students who answered the item correctly

R = represents the number of correct answers

N = the number of students taking the rest.

Thus, if 21 out of 26 students tested obtained the correct answer for one of the items, that item would have an index of difficulty (or a facility value) of 77 or 77 per cent.[45]

Calculation to obtain P, in order to analyze the degree of difficulty of 50 items in achievement test items followed by 72 people taste.

b. Item discrimination

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

Where :

D = Discriminatory power (item discrimination index numbers)

[45] J. B Heaton, Writing English Language Test, (London and New York), page

JA = the count of upper group

JB = the count lower group

BA= the count of upper group, who answered correctly

BB = the count of lower group, who answered correctly

PA $= \frac{BA}{JA}$ = proportion of participant who answered correctly the upper group

PB $= \frac{BB}{JB}$ = proportion participants who aanswered correctly the lower group

For calculating D

CHAPTER IV

RESERACH RESULT

A. Research Location

1. Brief History of SMPN 2 Jetis

The existing of SMPN 2 Jetis ponorogo was begun from the consciousness of local Government of Ngasinan and also the society around Ngasinan village to obtain learning opportunities for school age children of Ngasinan and surrounding village. Before the building of SMPN 2 Jetis was built and legalized from the Regional Ministry of Education and Culture. The Government programs to establish new schools immediately responded by the Local Government of Ponorogo regency and society, primarily the people in Ngasinan village by preparing the land or location of school construction and requirements as deemed necessary.

Later, on November 22th 1986, SMPN 2 Jetis was built and legalized by the Regional Ministry of Education and Culture. It is stated at Gajah Mada Street 13 Ngasinan village Jetis sub-district Ponorogo regency.

Since it was built in SMPN 2 Jetis has been happened commutation of headmaster. They are:[46]

a. Isran (1986-1993)

[46] Look at transcript of documentation number 02/D/19-V/2011

b. Suherman, B.A (1993-1999)

c. Hj. Siti Nurjanah, S.Pd (1999-2006)

d. Drs. Wahyu Hermadi (2006-2007)

e. Mulyono, S.Pd (2007-2010)

f. Drs. Dandun Santoso, M.Pd (2010-now)

2. The Geographical Location of SMPN 2 Jetis

SMPN 2 Jetis is located at:

Street : Gajah Mada

Number : 13

Village : Ngasinan

Sub-district : Jetis

Regency : Ponorogo

Province : East Java

SMPN 2 Jetis takes along 5.777 m^2 that is located 1,5 km in the south of Jetis intersection. It is a strategies place that is in the public transportation line.

3. The Vision and Mission of SMPN 2 Jetis Ponorogo

The vision of SMPN 2 Jetis is excellent in the academic achievement, qualified students and graduates, virtuous, and innovative.

To realize the vision, SMPN 2 Jetis create states its mission as follows:

a. Creating the academic achievement;

b. Achieving academic or non-academic tournament or competition;

c. Realizing the habitual of rules obedient;

d. Keeping the culture identity and virtuous character;

e. Realizing all positive activities in gaining the faithful and obedient to God Almighty;

f. Gaining the students' skill and ability of art in the form of wall paper making competition or other tournament.

The inherent purposes of education are laid in such aspects as intelligence, knowledge, personality, character, noble morality and great skills to live independently and to continue for higher education. Accordingly, the policies of ministry of education in SMPN 2 Jetis are expected to reach the following purposes:

a. Enlarging the students' potentials for their future needs.

b. Creating the students those are independent and discipline.

c. Having all the needed of educational facilities.

d. Realizing all the school programs.

e. Creating the school life in peace and harmony.

f. Achieving the graduates those have excellent quality, great intelligence, faithful and obedient to God Almighty.

B. Data Description

Data that have been analysis by the researcher is the answer sheet of the students. Its result of the try out test. There is multiple choice test that

contain 50 questions, each questions have 4 choice answer, namely A, B, C, D.

In this try out test, the reading section has four reading questions, starts from number 16 until 19, and then number 20 until 23, and number 28 until 32, and the last for reading test is number 35 until 39. For the greeting expression has two number, it is starts from number one and two.

For attention questions starts from number 3 and number 4. The type of the test of number 5-7 is giving messager. For the announcement questions starts from number 8 until 11. For the advertisement starts from number 12 until 15, number 33 until 34, number 43 until 45. For the procedure text in this test starts fror number 24 until 27. From number 40 until 42 is letter text. And for fill in the blank questions start from number 47 until 49. And the last for jumbled sentence is the number 50.

The student answer sheet take all answer sheets from the ninth grade as sample. And that the have analyzed their answer sheets, based in the answer key. After that the data have been analysis. The researcher analyzed the item difficulty, item discrimination, validity and reliability of the test.

After we try the test to the students, we get the result of the student's test. They are as follows :

Table 4.1
Distribucy Frequency Student's Score

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	26.00	1	1.4	1.4	1.4
	34.00	3	4.2	4.2	5.6
	36.00	3	4.2	4.2	9.9
	38.00	5	7.0	7.0	16.9
	40.00	1	1.4	1.4	18.3

	42.00	2	2.8	2.8	21.1
	44.00	4	5.6	5.6	26.8
	46.00	3	4.2	4.2	31.0
	48.00	4	5.6	5.6	36.6
	50.00	4	5.6	5.6	42.3
	52.00	2	2.8	2.8	45.1
	54.00	5	7.0	7.0	52.1
	56.00	2	2.8	2.8	54.9
	58.00	5	7.0	7.0	62.0
	60.00	6	8.5	8.5	70.4
	62.00	3	4.2	4.2	74.6
	64.00	1	1.4	1.4	76.1
	68.00	1	1.4	1.4	77.5
	70.00	2	2.8	2.8	80.3
	72.00	2	2.8	2.8	83.1
	74.00	3	4.2	4.2	87.3
	76.00	2	2.8	2.8	90.1
	78.00	2	2.8	2.8	93.0
	80.00	2	2.8	2.8	95.8
	82.00	3	4.2	4.2	100.0
	Total	72	100.0	100.0	

From the table distribusi frequency above, we can find there are 72 frequency.

Based on the table above, the histogram can be seen as follows:

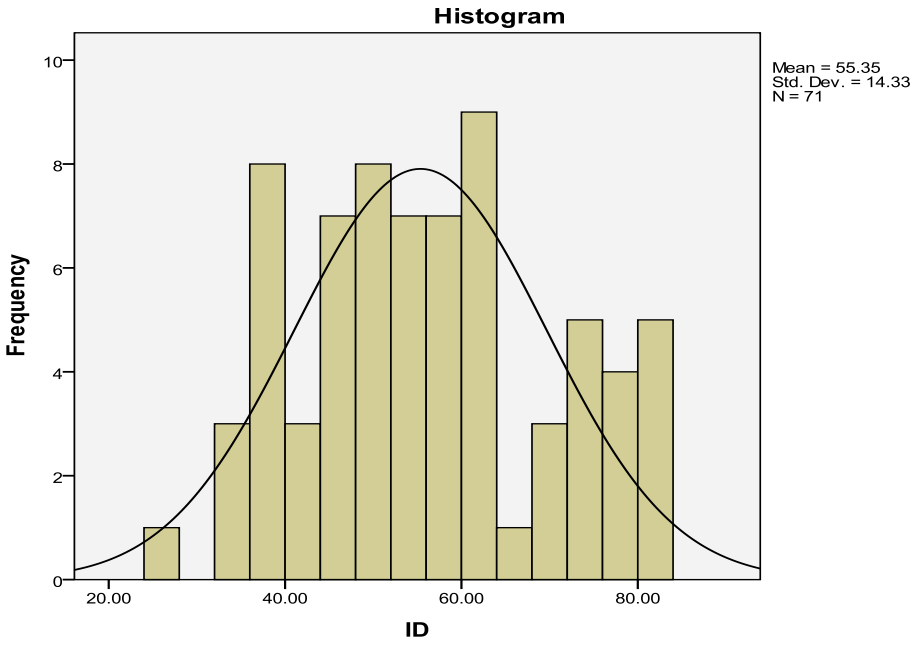

Figure:4.1 Histogram for Student's Score

From the histogram above, it is stated M = 55,35 and SD = 14,33. To determine the category student's score was good, medium, or low, the researcher grouped scores using the standart as follows:

1. More than M + 1.SD (55,35 + 14,33 = 67) is categorized into good
2. Between M – 1.SD to M + 1.SD (41 – 67) is categorized into medium
3. Less than M – 1.SD (55,35 – 14,33 = 41) is categorized into low

It can be seen that the scores which are more than 67 is considered good, while the scores which are less than 41 is low

C. Data Analysis

1. Validity

In every item (rxy) has the positive correlation more than 0,3 so the item is valid.[47] And if every item has the correlation less than 0,3 so the item is invalid. Finally, the result of the test validity is as follow:

Table 4.2
The Result of Validity Test

No Item	'r' arithmetic	'r' table	Explanation
1	0,307	0,30	**Valid**
2	0,352	0,30	**Valid**
3	0,153	0,30	**Invalid**
4	0,257	0,30	**Invalid**
5	0,224	0,30	**Invalid**
6	0,275	0,30	**Invalid**
7	0,377	0,30	**Valid**
8	0,229	0,30	**Invalid**
9	0,386	0,30	**Valid**
10	0,189	0,30	**Invalid**

[47] Sugiyono, Metode Penelitian Pendidikan: Pendekatan Kuantitatif, Kualitatif dan R&D,(Bandung:Alfabeta,2006),178.

11	0,338	0,30	**Valid**
12	0,380	0,30	**Valid**
13	0,279	0,30	**Invalid**
14	0,158	0,30	**Invalid**
15	0,160	0,30	**Invalid**
16	0,300	0,30	**Valid**
17	0,063	0,30	**Invalid**
18	0,201	0,30	**Invalid**
19	0,350	0,30	**Valid**
20	0,136	0,30	**Invalid**
21	0,227	0,30	**Invalid**
22	0,228	0,30	**Invalid**
23	0,208	0,30	**Invalid**
24	0,367	0,30	**Valid**
25	0,350	0,30	**Valid**
26	0,174	0,30	**Invalid**
27	0,264	0,30	**Invalid**
28	0,288	0,30	**Invalid**
29	0,308	0,30	**Valid**
30	0,337	0,30	**Valid**
31	0,294	0,30	**Invalid**
32	0,211	0,30	**Invalid**
33	0,429	0,30	**Valid**
34	0,300	0,30	**Valid**
35	0,306	0,30	**Valid**
36	0,125	0,30	**Invalid**
37	0,266	0,30	**Invalid**
38	0,346	0,30	**Valid**
39	0,170	0,30	**Invalid**
40	0,300	0,30	**Valid**
41	0,219	0,30	**Invalid**
42	0,148	0,30	**Invalid**
43	0,328	0,30	**Valid**
44	0,295	0,30	**Invalid**
45	0,399	0,30	**Valid**
46	0,372	0,30	**Valid**
47	0,251	0,30	**Invalid**
48	0,421	0,30	**Valid**
49	0,390	0,30	**Valid**
50	0,257	0,30	**Invalid**

From the table above, we can find 28 invalid numbers they are number

3,4,5,6,8,10,13,14,15,17,18,20,22,23,26,27,28,31,32,36,37,39,41,47,50.

And we find 22 valid numbers, they are number

1,2,7,9,11,12,16,19,24,25,29,30,33,34,35,38,40,43,45,46,48,49.

2. Reliabilty

In this research used Spearman Brown (Split half) formula to measure the reliability of the test. The formula is [48] :

$r_i = \frac{2rb}{1+rb}$

r_i = the internal reliability of all the instrument

r_b = the correlation of product moment between the first half and the second half

$r_i = \frac{2rb}{1+rb}$

$r_i = \frac{2\ x\ 0{,}76606}{1+0{,}76606}$

$r_i = \frac{1{,}532121}{1{,}76606} = 0{,}867536$

If $r_{count} > r_{table}$, the instrument is reliable

If $r_{count} > r_{table}$, the instrument is not reliable

Table 4.3
Recapitulation of Test Item Reliability

'r' arithmetic	**'r' table**	**Explanation**
0,867536	0,433	Reliable

From the interpretation above, this research has 21 of number. So, df = (21- 2) = 19. So,"r" table of 5% is 0,433, "r" count is 0,867536. It

[48] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif dan R&D,(Bandung:Alfabeta,2006),131.

can be concluded that ”r” count > ”r”table (0,867536> 0,433), so the instrument is reliable.

3. Item Difficulty

The difficulty of an item can be described statistically as the proportion of students who can answer the item correctly. The higher the value of difficulty, the easier the item.

Table 4.4
Calssification of Item Difficulty

The amount of P	**Interpretation**
Less than 0,30	To difficult
0,30-0,70	Medium
More than 0,70	Too easy

The result of analysis classified item difficulty in three groups :

1. P> 0.7 (easy) = if prop. Correct higher than 0.7 this item classified as easy item.
2. $0.3 \leq p \leq 0.70$ (medium) = if prop. Correct hihger than or as same as 0.3 and lower than or as same as 0.7 this item classified as moderate item.
3. P < 0.3 (difficult) = if prop. Correct lower than 0.3 this item classified as difficult item

Table 4.5

Analysis prop. Correct/difficulty

Item number	Item difficulty (P) $P = \frac{N \quad p}{N}$	Interpretation
1	$P = \frac{52}{72} = 0{,}72$	Too easy
2	$P = \frac{54}{72} = 0{,}75$	Too easy
3	$P = \frac{46}{72} = 0{,}63$	Medium
4	$P = \frac{44}{72} = 0{,}61$	Medium
5	$P = \frac{55}{72} = 0{,}76$	Too easy
6	$P = \frac{49}{72} = 0{,}68$	Medium
7	$P = \frac{55}{72} = 0{,}76$	Too easy
8	$P = \frac{47}{72} = 0{,}65$	Medium
9	$P = \frac{38}{72} = 0{,}52$	Medium
10	$P = \frac{46}{72} = 0{,}63$	Medium
11	$P = \frac{51}{72} = 0{,}70$	Too easy
12	$P = \frac{39}{72} = 0{,}54$	Medium
13	$P = \frac{44}{72} = 0{,}61$	Medium
14	$P = \frac{52}{72} = 0{,}72$	Too easy
15	$P = \frac{44}{72} = 0{,}61$	Medium
16	$P = \frac{53}{72} = 0{,}73$	Too easy
17	$P = \frac{58}{72} = 0{,}80$	Too easy
18	$P = \frac{48}{72} = 0{,}66$	Medium

19	$P = \frac{45}{72} = 0,62$	Medium
20	$P = \frac{40}{72} = 0,55$	Medium
21	$P = \frac{43}{72} = 0,59$	Medium
22	$P = \frac{48}{72} = 0,66$	Medium
23	$P = \frac{41}{72} = 0,56$	Medium
24	$P = \frac{51}{72} = 0,70$	Too easy
25	$P = \frac{52}{72} = 0,72$	Too easy
26	$P = \frac{43}{72} = 0,59$	Medium
27	$P = \frac{52}{72} = 0,72$	Too easy
28	$P = \frac{52}{72} = 0,72$	Too easy
29	$P = \frac{36}{72} = 0,5$	Medium
30	$P = \frac{47}{72} = 0,65$	Medium
31	$P = \frac{39}{72} = 0,54$	Medium
32	$P = \frac{39}{72} = 0,54$	Medium
33	$P = \frac{51}{72} = 0,70$	Too easy
34	$P = \frac{38}{72} = 0,52$	Medium
35	$P = \frac{47}{72} = 0,65$	Medium
36	$P = \frac{35}{72} = 0,48$	Medium
37	$P = \frac{57}{72} = 0,79$	Too easy
38	$P = \frac{50}{72} = 0,69$	Medium
39	$P = \frac{47}{72} = 0,65$	Medium
40	$P = \frac{35}{72} = 0,48$	Medium

41	$P = \frac{49}{72} = 0{,}68$	Medium
42	$P = \frac{55}{72} = 0{,}76$	Too easy
43	$P = \frac{44}{72} = 0{,}61$	Medium
44	$P = \frac{41}{72} = 0{,}56$	Medium
45	$P = \frac{45}{72} = 0{,}62$	Medium
46	$P = \frac{54}{72} = 0{,}75$	Too easy
47	$P = \frac{46}{72} = 0{,}63$	Medium
48	$P = \frac{45}{72} = 0{,}62$	Medium
49	$P = \frac{45}{72} = 0{,}62$	Medium
50	$P = \frac{56}{72} = 0{,}77$	Too easy

From the table above, there are three classification in item difficulty level. They are as follows :

Table 4.6
Classification of Item Difficulty

Classification	No. Items	Percent (%)
Too easy	1,2,5,7,11,14,16,17,24,25,27,28,33, 37,42,46,50	34 %
Medium	2,4,6,8,9,10,12,13,15,18,19,20,21, 22,23,26,29,30,31,32,34,35,36,38, 39,40,41,43,44,45,47,48,49	66 %
Too diificult	-	0 %

Based on that classification above, the easy item showed 34% too easy item, 66% medium item, and 0% difficult item.From table 4.6 this consist many too easy item. In addition from table 4.6 showed item 7 has prop correct 0,76 it means is too easy.

From table 4.7 showed column number of the student, the column give information how many students can answer correctly that item. From table showed there are 52 students can answer correctly from item no. 1 from 72 students, it means 20 students who answer incorrectly. So, this item classified as easy item. From item no 2. There are 54 students who answer correctly and this item classsified has medium item. And for items classified difficulty item is empty.

4. Item Discrimination

A good item should be able to discriminate students with high scores from those low scores. To score item discrimintion can be classified as below:

Table 4.7
Classification of Item Discrimination

The amount of indeks diskriminasi item (D)	Classified :	Interpretation :
Less than 0,20	Bad	An items in question once the distinguishing weak is not considered a good distinguishing features
0,20 – 0,40	Fairly bad	An items term in question has enough distinguishing features (being)
0,40 – 0,70	Fairly good	An item in question already has a fairly good differentiator
0,70 – 1,00	Good	An item in question already has an good differentiator.
Are negative	Very bad	Item in question different negative

		power

They are five classification in item discrimination. They are bad, fairly bad, fairly good, good, very bad. Before we classify the result of item, we should count/examine the upper and lower group.

To know the discrimination, we should count it using the formula :

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

The explanation as follows :

Table 4.8
An Analysis of Item Discrimination

NO	PA= $\frac{BA}{JA}$	PB = $\frac{BB}{JB}$
1	PA = $\frac{23}{26} = 0{,}88$	PB = $\frac{29}{46} = 0{,}63$
2	PA = $\frac{23}{26} = 0{,}88$	PB = $\frac{31}{46} = 0{,}67$
3	PA = $\frac{17}{26} = 0{,}65$	PB = $\frac{29}{46} = 0{,}63$
4	PA = $\frac{20}{26} = 0{,}77$	PB = $\frac{24}{46} = 0{,}53$
5	PA = $\frac{22}{26} = 0{,}85$	PB = $\frac{33}{46} = 0{,}72$
6	PA = $\frac{23}{26} = 0{,}88$	PB = $\frac{26}{46} = 0{,}63$
7	PA = $\frac{25}{26} = 0.96$	PB = $\frac{30}{46} = 0{,}65$
8	PA = $\frac{19}{26} = 0{,}73$	PB = $\frac{28}{46} = 0{,}61$
9	PA = $\frac{20}{26} = 0{,}77$	PB = $\frac{18}{46} = 0{,}39$
10	PA = $\frac{19}{26} = 0{,}73$	PB = $\frac{27}{46} = 0{,}59$
11	PA = $\frac{19}{26} = 0{,}73$	PB = $\frac{27}{46} = 0{,}59$
12	PA = $\frac{20}{26} = 0{,}77$	PB = $\frac{19}{46} = 0{,}41$
13	PA = $\frac{20}{26} = 0{,}77$	PB = $\frac{24}{46} = 0{,}52$
14	PA = $\frac{21}{26} = 0{,}81$	PB = $\frac{31}{46} = 0{,}67$
15	PA = $\frac{18}{26} = 0{,}69$	PB = $\frac{26}{46} = 0{,}63$

16	PA $= \frac{23}{26} = 0{,}88$	PB $= \frac{30}{46} = 0{,}65$
17	PA $= \frac{22}{26} = 0{,}85$	PB $= \frac{36}{46} = 0{,}78$
18	PA $= \frac{21}{26} = 0{,}81$	PB $= \frac{27}{46} = 0{,}59$
19	PA $= \frac{21}{26} = 0{,}81$	PB $= \frac{24}{46} = 0{,}52$
20	PA $= \frac{15}{26} = 0{,}58$	PB $= \frac{25}{46} = 0{,}54$
21	PA $= \frac{18}{26} = 0{,}69$	PB $= \frac{25}{46} = 0{,}54$
22	PA $= \frac{19}{26} = 0{,}73$	PB $= \frac{29}{46} = 0{,}63$
23	PA $= \frac{18}{26} = 0{,}69$	PB $= \frac{23}{46} = 0{,}5$
24	PA $= \frac{25}{26} = 0{,}96$	PB $= \frac{26}{46} = 0{,}63$
25	PA $= \frac{24}{26} = 0{,}92$	PB $= \frac{28}{46} = 0{,}61$
26	PA $= \frac{18}{26} = 0{,}69$	PB $= \frac{25}{46} = 0{,}54$
27	PA $= \frac{21}{26} = 0{,}81$	PB $= \frac{31}{46} = 0{,}67$
28	PA $= \frac{23}{26} = 0{,}88$	PB $= \frac{29}{46} = 0{,}63$
29	PA $= \frac{18}{26} = 0{,}69$	PB $= \frac{18}{46} = 0{,}39$
30	PA $= \frac{21}{26} = 0{,}81$	PB $= \frac{26}{46} = 0{,}63$
31	PA $= \frac{19}{26} = 0{,}73$	PB $= \frac{20}{46} = 0{,}43$
32	PA $= \frac{17}{26} = 0{,}65$	PB $= \frac{22}{46} = 0{,}48$
33	PA $= \frac{25}{26} = 0{,}96$	PB $= \frac{26}{46} = 0{,}63$
34	PA $= \frac{19}{26} = 0{,}73$	PB $= \frac{19}{46} = 0{,}41$
35	PA $= \frac{22}{26} = 0{,}85$	PB $= \frac{25}{46} = 0{,}54$
36	PA $= \frac{15}{26} = 0{,}58$	PB $= \frac{20}{46} = 0{,}43$
37	PA $= \frac{24}{26} = 0{,}92$	PB $= \frac{33}{46} = 0{,}72$
38	PA $= \frac{22}{26} = 0{,}85$	PB $= \frac{28}{46} = 0{,}61$
39	PA $= \frac{20}{26} = 0{,}77$	PB $= \frac{27}{46} = 0{,}59$
40	PA $= \frac{18}{26} = 0{,}69$	PB $= \frac{17}{46} = 0{,}37$
41	PA $= \frac{21}{26} = 0{,}81$	PB $= \frac{28}{46} = 0{,}61$

42	PA $= \frac{23}{26} = 0{,}88$	PB $= \frac{32}{46} = 0{,}69$
43	PA $= \frac{20}{26} = 0{,}77$	PB $= \frac{24}{46} = 0{,}52$
44	PA $= \frac{20}{26} = 0{,}77$	PB $= \frac{21}{46} = 0{,}46$
45	PA $= \frac{22}{26} = 0{,}85$	PB $= \frac{23}{46} = 0{,}5$
46	PA $= \frac{26}{26} = 1$	PB $= \frac{28}{46} = 0{,}61$
47	PA $= \frac{21}{26} = 0{,}88$	PB $= \frac{25}{46} = 0{,}54$
48	PA $= \frac{48}{26} = 0{,}88$	PB $= \frac{22}{46} = 0{,}48$
49	PA $= \frac{22}{26} = 0{,}85$	PB $= \frac{23}{46} = 0{,}5$
50	PA $= \frac{23}{26} = 0{,}88$	PB $= \frac{33}{46} = 0{,}72$

To know the discrimination, we should count it using the formula :

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

The explanation as follows :

Table 4.9
An Analysis of Item Discrimination

No	P = correct answers / count of students D = PA – PB	Interpretation
1	D = 0,88 – 0,63 = 0,25	**Fairly bad**
2	D = 0,88 – 0,67 = 0,21	**Fairly bad**
3	D = 0,65 – 0,63 = 0,02	**Bad**
4	D = 0,77 – 0,52 = 0,25	**Fairly bad**
5	D = 0,85 – 0,72 = 0,13	**Bad**
6	D = 0,88 – 0,63 = 0,25	**Fairly bad**
7	D = 0,96 – 0,65 = 0,31	**Fairly bad**
8	D = 0,73 – 0,61 = 0,12	**Bad**
9	D = 0,77 – 0,39 = 0,38	**Fairly bad**
10	D = 0,73 – 0,59 = 0,14	**Bad**
11	D = 0,92 – 0,59 = 0,33	**Fairly bad**
12	D = 0,77 – 0,41 = 0,36	**Fairly bad**
13	D = 0,77 – 0,52 = 0,25	**Fairly bad**
14	D = 0,81 – 0,67 = 0,14	**Bad**
15	D = 0,69 – 0,63 = 0,06	**Bad**

16	D = 0,88 – 0,65 = 0,23	**Fairly bad**
17	D = 0,85 – 0,78 = 0,07	**Bad**
18	D = 0,81 – 0,59 = 0,22	**Fairly bad**
19	D = 0,81 – 0,52 = 0,29	**Fairly bad**
20	D = 0,58 – 0,54 = 0,04	**Bad**
21	D = 0,69 – 0,54 = 0,15	**Bad**
22	D = 0,73 – 0,63 = 0,1	**Bad**
23	D = 0,69 – 0,5 = 0,19	**Bad**
24	D = 0,96 – 0,63 = 0,33	**Fairly bad**
25	D = 0,92 – 0,61 = 0,31	**Fairly bad**
26	D = 0,69 – 0,54 = 0,15	**Bad**
27	D = 0,81 – 0,67 = 0,14	**Bad**
28	D = 0,88 – 0,63 = 0,25	**Fairly bad**
29	D = 0,69 – 0,39 = 0,3	**Bad**
30	D = 0,81 – 0,63 = 0,18	**Bad**
31	D = 0,73 – 0,43 = 0,3	**Bad**
32	D = 0,65 – 0,48 = 0,17	**Bad**
33	D = 0,96 – 0,63 = 0,33	**Fairly bad**
34	D = 0,73 – 0,41 = 0,32	**Fairly bad**
35	D = 0,85 – 0,54 = 0,31	**Fairly bad**
36	D = 0,58 – 0,43 = 0,43	**Bad**
37	D = 0,92 – 0,72 = 0,2	**Bad**
38	D = 0,85 – 0,61 = 0,24	**Fairly bad**
39	D = 0,77 – 0,59 = 0,18	**Bad**
40	D = 0,69 – 0,37 = 0,32	**Fairly bad**
41	D = 0,81 – 0,61 = 0,2	**Bad**
42	D = 0,88 – 0,69 = 0,19	**Bad**
43	D = 0,77 – 0,52 = 0,25	**Fairly bad**
44	D = 0,77 – 0,46 = 0,31	**Fairly bad**
45	D = 0,85 – 0,5 = 0,35	**Fairly bad**
46	D = 1 – 0,61 = 0,39	**Fairly bad**
47	D = 0,88 – 0,54 = 0,34	**Fairly bad**
48	D = 0,88 – 0,48 = 0,4	**Bad**
49	D = 0,85 – 0,5 = 0,35	**Fairly bad**
50	D = 0,88 – 0,72 = 0,16	**Bad**

Table 4.10
Classification of Item Discrimination

Classification	No. Items	Percent (%)

Bad	3,5,8,10,14,15,17,20,21,22,23,26,27,29, 30,31,32,36,37,39,41,42,48,50	48 %
Fairly bad	1,2,4,6,7,9,11,12,13,16,18,19,24,25,28, 33,34,35,38,40,43,44,45,46,47,49	52 %
Fairly good	-	0%
Good	-	0%
Very bad	-	0%

The result of the analysis in table 4.10 showed that the test consist of 52% items has fairly bad item discrimination, 48% items has bad item discrimination.

From table 4.10 showed percentage of fairly bad item dscrimination is high namely 52%. And this table also showed percentage of bad items lower then fairly bad items.

D. Discussion and Interpretation

The purpose of this study was giving the picture of English Try Out test in SMP N 2 Jetis by presenting An Analysis of multiple choice item on English Try Out test administered to the student of ninth grade of SMP N 2 Jetis.

Additionally, the try out test needs to be analyzed, to know the validity and reliability of the items. Some items can be use to the next try out test, and some of them should be revised, but for some items it's useless.

From the explaining above can concluded that the test has low validity, because there are many item medium, many items classified as bad item discrimination.

Reffering the explaining above, the validity of item test thay analysis with item difficulty, and item discrimination. The try out test consist many

medium items. It is draw from explaining above that 66% from total item classified as medium item. This condition show that most students can answer correctly.

Based on the analysis above this test has low reliability. Its showed from the scale Alpha = 0,86. From the result showed that, the test that has low validity it's that the test also has low reliability.

Analysis item discrimination showed that consist of 48% item has bad item discrimination, 52% item has fairly bad item discrimination. It is means that there are many items can't differentiate the upper and lower group students.

Finally , we as teachers are well as test makers must analyzed teh result of the exam and many feedback we many have obtained from our students and teachers. it is also vitally important to reiterate that there is still muct to be learned about language testing, such as the issues of reliability, validity and item analysis all of which were mentioned in this research. However, if we are open to the possibility of constant revision of our language exams, then we can surely be more succesful in creating exams that are more effective.

CHAPTER V

CLOSING

A. Conclusion

1. Multipe choice test has low validity. It was drawn from the result of analysis 22 items classified as valid criteria. And 28 items classified as invalid criteria. Its mean that the validity of the test is low.
2. Reliability of multiple choice test classified low reliability. The scale Alpha showed reliability of the test with 0,86. From the result showed that it is reliability.
3. This test has medium difficulty items, the easy item showed 34% too easy item, 66% medium item, and 0% difficult item. In which, there are 31 items from 50 items have medium criteria. So this test is medium test for student, which some of students can answer the questions.
4. This test has fairly bad discrimination items. It is drawn from the results of analysis that showes 52% items is classified as fairly bad items. Where there are 26 from 50 items that includes fairly bad criteria. Then, 48% is classified as bad item.

B. Recomendation

Based on the result of the research, the researcher would like offer some recomendation to the teacher who made the try out test need analyzed the item tests, for the result more good for the next try out test, beacause the try out test is important to try the test before the final test. But in item difficulty level has medium item, it's means that the students can answer correctly. In this case,

who made the try out test item need to increase the reliability of the test. For the teacher or who made the test needs too increase the validity the items.