ITEM ANALYSIS OF TRY-OUT TEST (COMPUTER-BASED TEST) AT MTS N 2 PONOROGO

THESIS

By

FRENDY EKO WIDYANTORO

NIM. 210915047

ENGLISH EDUCATION DEPARTMENT

FACULTY OF TARBIYAH AND TEACHER TRAINING

STATE INSTITUTE OF ISLAMIC STUDIES PONOROGO

JUNE 2019

ABSTRACT

WIDYANTORO, FRENDY EKO. 2019. Item Analysis of Try-out Test (Computer-Based Test) at MTs N 2 Ponorogo. Thesis, English Education Department, Tarbiyah Faculty, State Institute of Islamic Studies of Ponorogo. Advisor: Ahmad Nadhif, M. Pd.

Keyword: Evaluation, Assessment, Measurement, Test Item Analysis.

The purposes of this study were to determine the degree of validity, reliability, item difficulty, item discrimination, and item distractor of the test.

This research applied quantitative approach with discriptive research design. The subjects in this research were the try-out (Computer-Based Test) at MTs N 2 Ponorogo which consisted of 50 questions. The procedure of data collection was documentation. The data were analyzed using product-moment formula for validity, split-half formula for reliability, IF= $\frac{N_{correct}}{N_{total}}$ formula for item difficulty, ID = $IF_{upper} + IF_{lower}$ formula for item discrimination, and $p = \frac{n_{swd}}{n_{ts}} \times 100\%$ formula for distractor item which were calculated by IBM SPSS program version 23.

The finding revealed that the test had low validity, with r_{table} 0,113 and significance level 5%. It was drawn from the result of analysis that 4% of the item test had very low validity, 46% had low validity, 38% had medium validity, and 4% had high validity. The finding revealed that the test had high reliability by alpha scale 0,867 and r_{table} was 0,113. The finding showed that the level of difficulty of the test was difficult. It was proven by the result of the analysis that there were 32% easy items, 22% medium items, and 46% difficult items. The finding revealed that the test had very good item discrimination where 21 of 50 items had very good item discrimination, there were 4% very poor items, 10% poor items, 22% medium items, 22% good items, and 42% very good items. And, from the result of test analysis revealed that the test was good in item distractor, there were 9 questions that some items needed to be revised.

APPROVAL SHEET

This is to certify that Sarjana's thesis of:

Name	: Frendy Eko Widyantoro
Student Number	: 210915047
Faculty	: Tarbiyah and Teacher Training
Departement	: English Education
Title	: Item Analysis of Try-out Test (Computer-Based Test) at MTs N 2 Ponorogo

Has been approved by the advisor and is recommended for approval and acceptance.

Advisor

Ahmad Nadhif, M. Pd. Ponorogo, June 26th 2019
NIP. 198004182008011009

Acknowledged by
Head of English Education Department of
Tarbiyah and Teachers Training Faculty
State Institute of Islamic Studies Ponorogo

Prvia Rochmahwati, M. Pd.
NIP. 198103162011012003

MINISTRY OF RELIGIOUS AFFAIRS

STATE INSTITUTE OF ISLAMIC STUDIES PONOROGO

RATIFICATION

This is to certify that Sarjana's thesis of:

Name	: Frendy Eko Widyantoro
Student Number	: 210915047
Faculty	: Tarbiyah and Teacher Training
Department	: English Education
Title	: Item Analysis of Try-out Test (Computer-Based Test) at MTs N 2 Ponorogo

Has been approved by the board of examiners on:

Day	: Monday
Date	: July 22nd, 2019

And has been accepted as the requirement for the degree of the sarjana in English Education on:

Day	: Tuesday
Date	: July 23rd, 2019

Ponorogo, July 23rd, 2019

Certified by

Dean of Tarbiyah and Teachers Training

State Institute of Islamic Studies

Ponorogo

Dr. Ahmadi, M. Ag.

NIP. 196512171997031003

Board of Examiners

1. Chairman : Dr. Ahmadi, M. Ag ( )
2. Examiner I : Dra. Aries Fitriani, M. Pd ( )
3. Examiner II : Ahmad Nadhif, M. Pd ( )

SURAT PERSETUJUAN PUBLIKASI

Yang bertanda tangan di bawah ini:

Nama	: FRENDY EKO WIDYANTORO
NIM	: 210915047
Fakultas	: Tarbiyah dan Ilmu Keguruan
Program Studi	: Tadris Bahasa Inggris
Judul Skripsi/Tesis	: ITEM ANALYSIS OF TRY-OUT TEST (COMPUTER-BASED TEST) AT MTS N 2 PONOROGO

Menyatakan bahwa naskah skripsi/tesis telah diperiksa dan disahkan oleh dosen pembimbing. Selanjutnya saya bersedia naskah tersebut dipublikasikan oleh perpustakaan IAIN Ponorogo yang dapat diakses di ethesis.iainponorogo.ac.id. Adapun isi dari keseluruhan tulisan tersebut, sepenuhnya menjadi tanggung jawab dari penulis.

Demikian pernyataan saya untuk dapat dipergunaan semestinya.

Ponorogo, 26 Juli 2019

Penulis

Frendy Eko Widyantoro

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	:	FRENDY EKO WIDYANTORO
NIM	:	210915047
Jurusan	:	Tadris Bahasa Inggris
Fakultas	:	Tarbiyah dan Ilmu Keguruan
Judul Skripsi	:	Item Analysis of Try-out Test (Computer-Based Test) at MTs N 2 Ponorogo

Dengan ini, menyatakan dengan sebenarnya bahwa skripsi yang saya tulis ini adalah benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan tulisan atau pikiran orang lain yang saya aku sebagai hasil tulisan atau pikiran saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Ponorogo, 26 Juni 2019

Yang membuat pernyataan

:ndy Eko Widyantoro

NIM. 210915047

CHAPTER I

INTRODUCTION

In this chapter, the researcher discusses about the background of the study, limitation of the problem, statement of the problem, objective of the study, significances of the study, and organization of the study.

A. Background of the Study

Evaluation, assessment, measurement, and test cannot be separated from the education system. That is because all of them relate to each other. Evaluation is a process of collecting data to determine how far, in terms of what, and in which part the educational goals have been achieved. Evaluation is also defined as the systematic collection and analysis of all relevant information necessary to promote the improvement of the curriculum and analyze its effectiveness within the

context of the particular institutions.[1] Evaluation is used to determine the teachers' level of success in the learning process and measure how far the learning objectives have been achieved, and can also be used to make decisions in improving learning. Evaluation as one of the main components of the learning process should be understood, planned and implemented in an effort to support the success of improving the learning process.

The first thing that must be conducted in evaluation is collecting data. This process is the process of collecting data about the program to be evaluated so that the evaluation is closely related to the use of measurement and data collection instruments. In the process of collecing data, data must be ensured to be

[1] James Dean Brown. *Testing in Language Programs,* (New Jersey: Prentice Halls, 1996), 277.

truly valid or good. Without good data then evaluation results will result in interpretation mistakes. When the teacher is conducting an evaluation, the main process is gathering as much information as possible to assist the process of granting the learning score. After the data collection process, the data is analyzed and designed for the evaluation program.

Evaluation can be conducted after it is based on a form of assessment. Assessment is the process of collecting, synthesizing, and interpreting information to aid in decision making.[2] Therefore, assessment is conducted to gather the information from the students about what they know and what they can do. After the process of gathering information, the teacher can

[2] Michael K. Russell and Peter W. Airasian. *Classroom Assessment: Concept and Application*. (New York: McGraw Hill, 2012), 3.

arrange the assessment in a structured and systematic manner. The assessment is contained of measurement.

Measurement is process of quantifying or assigning a number to a performance or trait.[3] Measurement will give the answer the question about how much. As a result, the measurements will present quantitative results in the form of numbers or score. Measurement can be arranged in the form of test or non-test. Test is divided into three which are: based on how to do test, based on how to answer test, and based on Bloom's taxonomy. Non-test is categorized into: performance assessment, portfolio, project assessment, product assessment, self-assessment, peer-assessment, and attitude assessment. This study analyzes the types of assessments in the form of tests.

[3] Ibid, 11.

Test is defined as a way to organize evaluation in the form of tasks that must be done by the participant of the test.[4] Test is one of way to interpret ability of the students and also interpret the quality of the test. Teacher will know the quality of the test by analyzing the result of the test. The result of the test will show the quality of the test, with good or bad results. If the result of the test is good, the quality of the test can be said as good. Therefore, the test just need a little improvements. In contrast, if the test shows poor result, the test needs some revisions in some aspecs. The revision can be conducted by changing question grade or item distractors.

In evaluation, a test can be said to be good if it has validity, reliability, authenticity, practicability, and

[4] Sri Wahyuni and Abd. Syukur Ibrahim. *Asesmen Pembelajaran Bahasa*. (Bandung: PT Refika Aditama, 2012), 11.

washback.[5] The test can be said as good in validity, if it can show valid result in measuring result of the test or it has high validity. Then, the test can be said reliable, if the result of the test can be considered consistent or stable or it has high reliability. The next is practically, it means that the test has good procedure, efficient, and cheap. Therefore, the test has been structured before conducted. Then, authenticity of the test refers to the contents of the test. The contents of the test must be authentic or have natural language, relevant topic, and tasks represent. Washback refers to the effect of the test have on instructions in terms how of the students prepare for the test.[6]

[5] H. Douglas Brown. *Language Assessment Principles and Classroom Practices,* (New York: Longman, 2000), 19.

[6] Ibid, 28.

The quality of the test can be determined by conducting test items analysis. A test item analysis is the systematic evaluation of the effectiveness of the individual items on a test.[7] The test item analysis aims to identify either the question is a good, less good or not a good question. Test items test is also used to measure the skill of the students. In addition, test items analysis is used to get the feedback from the students and also to give the feedback for the students. After, test makers know the quality of the test or get the result of the test, the result must be revised. Therefore, the test is actually made up of the good quality questions. Then, a good question can be kept or improved. The activity of revised the test items is conducted in order to make the

[7] James Dean Brown. *Testing in Language Programs: A Comprehensive Guide to English Language Assessment,* (New York: The Mac-Graw Hills Companies, 2005), 41.

test qualified enough to be used as a measurement tool of student's learning outcomes. The test item analysis is done by calculating the aspect of validity, reliability, level of difficulty, discrimination index, and distribution pattern answer.

The test item analysis will reveal the degree of validity, reliability, level of difficulty, item discrimination, and item distractor.

From the explanation above, the researcher want to conduct item analysis of Try-Out tests (Computer-Based Test) in MTs N 2 Ponorogo. The researcher want to analyze the test at MTs N 2 Ponorogo using IBM SPSS. In this school, the test has never been analyzed before.[8]

B. Limitation of the Problem

[8] Interviewed at MTs N 2 Ponorogo on, March 2nd 2019.

Based on the background of the study outlined above, this research is restricted to look for the quality of the questions of the try-out (Computer-Based Test) in English subject in IX grade MTs Negeri 2 Ponorogo academic year of 2018/2019 in term of reliability, validity, item difficulty, item discrimination, and item distractor.

C. Statement of the Problem

From limitation of the problem the researcher sum that will find:

1. How is the try-out test fulfill the criteria of the good test in terms of validity degree?
2. How is the try-out test fulfill the criteria of the good test in terms of reliability degree?
3. How is the try-out test fulfill the criteria of the good test in terms of item difficulty degree?

4. How is the try-out test fulfill the criteria of the good test in terms of item discrimination degree?
5. How is the try-out test fulfill the criteria of the good test in terms of item distractor degree?

D. Objective of the Study

Based on the formulation of the problem above, the purpose of this study is to determine the test quality of the questions of the try-out questions in English subject in IX grade MTs Negeri 2 Ponorogo academic year of 2018/2019.

E. Significance of the Study

The researcher hopes that this research will be useful for teacher, and the readers. The result can be found as follows:

1. Theoretically

a. The result of this study will be useful and contribute to scientific treasure in the field of education.

b. For the purpose of scientific study and for information and reference for other researchers who want conduct further research.

2. Practically

a. For the teachers

This study provides suggestions for English teacher in particular about test item analysis and encourages teachers to be able to conduct the test item analysis in question which is used to improve the quality of the tests.

b. For the readers

This study is expected to give contribution for the readers in improving knowledge, particularly the student in IAIN Ponorogo.

c. For the researcher

This research is expected to be used by researcher as a provision in the future, if the researcher becomes an educator in the future, applying the knowledge gained in college and adding experience.

F. Organization of the Study

To make easier in writing the thesis, the thesis is divided into five chapters as follows:

The first chapter is introduction which contains of background of the study, statement of the problem,

objective of the study, significance of the study, limitations of the problems, and organization of thesis.

The second chapter is review of related literature, theoretical background, previous study, and theoretical framework which contains of review of related literature, evaluation, assessment, test, criteria of the good test, test item analysis, try-out (computer-based test), and theoretical framework.

The third chapter is research methodology which contains of research design, instrument of data collection, techniques of data collection, and technique of data analysis.

The fourth chapter is research result which explains about research location, description, analysis of the data and the interpretation of the finding result.

The fifth chapter is closing which contains about the conclusion and suggestion.

CHAPTER II

PREVIOUS RESEARCH FINDING, THEORITICAL BACKGROUND, THEORITICAL FRAMEWORK

In this chapter, the researcher discusses about previous research finding, theoretical background, and theoretical framework.

A. Previous Research Finding

This study discusses about item analysis of try-out test (Computer-Based Test). There are some relevant researches related to this research. First, similar research is thesis from Noorrachma Chandra Novianti entitled ”Test Item Analysis of the Final Examination on Economics Subject in Grade XII IPS SMA Negeri 1 Wonosari Academic Year 2014/2015”. The researcher analyzes about validity, reliability, level of difficulty, discrimination index, and distribution pattern answer

using program ANATES version 4.0.9. The result of the study, the researcher concludes that in terms of validity is good because the 33 items (82.5%) including to the valid questions, in terms of reliability is high or reliable because it has a high reliability which is equal to 0.87, in terms of level of difficulty is not good because the 32 items (80,%) include to easy category, in terms of discrimination Index is good because the only 11 items (27.5%) include to poor category, in terms of distribution pattern answer has a fairly good functioning distractors because there are 6 items (15%) has less good functioning distractors, and 9 items (22.5 %) have not good functioning distractors, and based on the overall analysis of the validity, level of difficulty, discrimination index and distribution pattern answers is less good because there are 8 items (20%) has a good

quality, 16 items (40%) has a less good quality, and 16 items (40%) has a not good quality.

Second, similar reseach is journal from Anita Noveria entitled ”Item Analysis on the Validity and the Reliability of the English Summative Test for the first-year students at MA Madani Alauddin Pao-pao”. The researcher analyzes about validity and reliability. The result of the study shows that the validity is different for the two kinds of test. First, the short-answer test is invalid as 2 out of 10 items (20%) are unable to deal with the standard of validity index required by at trustworthy test item; the ability of this item to measure what is supposed to measure. On the contrary, 8 out of 10 items (80%) are valid for they show the validity standard of a good test. Besides, the completion test is valid as 5 out of 5 items (100%) are able to deal with

the standard of validity index and the result is higher than critical value of product moment. Second, the reliability did not show the same result. The short-answer test is found good and trustworthy because the reliability index is 0.808 which is higher than the table of critical value of product moment (0.297) with the level of significance 95 %. However, the reliability of completion test is found to be not reliable as the reliability index is 0.140 which is lower than the table of critical value of product moment (0.297) with the level of significance 95 %.

Third, similar research is journal from Sefik Yasar and Asli Gundogan Cogenli entitled ”Determining Validity and Reliability of Data Gathering Instruments used by Program Evaluation Studies in Turkey”. The researcher determines validity

and reliability of data gathering instruments used by program evaluation studies. The result shows that the program evaluation studies conducted in Turkey, measurement tool development principles are not adequately adhered to, validity studies are usually limited to specialist's opinion, reliability studies however are mostly in the shape of analyzing the internal consistency and in the sense of determination reliability studies cannot be adequately made.

Fourth, similar reseach is thesis from Rafikasari Risqi Sakti entitled "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". The researcher analyzes about reliability, validity, item difficulty, and item descrimination. The result shows that multiple choice test has low validity

because from the result of analysis 22 items classified as valid criteria and 28 items classified as invalid criteria, in term of reliability is low because the scale of alpha shows realibility of the test with 0,86; in term of difficulty is medium because 34% of the items shows easy item, 66% medium item, and 0% difficult item; in term of descrimination is bad beacuse 52% of the items shows fairly bad items and 48% clasified as bad item.

Fifth, similar reseach is journal from Sibel Toksöz and Ayşe Ertunç entitled"Item Analysis of a Multiple-Choice Exam". The researcher analyzes item discrimination, item facility, and distractor efficiency using IBM SPSS version 20. The result shows that the multiple choice items seemed to be efficient in terms of item facility. Most of the items have acceptable item facility indexes which mean that the difficulty levels of

the items are suitable for the students. On the other hand, although the responses seem to be distributed evenly, there are some responses which are not discriminating enough between high and low ability students. These items are found to need revision to improve the discriminatory power and the quality of the exam overall. By doing so, the potential negative washback effect of the exam for high ability students could be diminished or even inhibited. The results of the analysis have also identified that there are some distractors which seemed to be completely inefficient.

The different this research between research above is this study analyzes try-out test which is conducted based on computer-based. This resesarch analyzes the test using computer program IBM SPSS version 23, which determines the degree of validity,

reliability, item discrimination, item diffiulty, and item distractor.

B. Theoretical Background

Theoretical background discusses about the subject of the study which are: evaluation, assessment, test, criteria of the good test, test item analysis, and computer-based test. The explanations are as follows:

1. Evaluation

Evaluation is the process of judging the quality or value of the performance or a course of action.[9] According to Brown, evaluation is defined as the systematic collection and analysis of all relevant information necessary to promote the improvement of the curriculum and analyze its

[9] Russell and Airasian, *Classroom Assessment: Concept and Application*, 10.

effectiveness within the context of the particular institutions.[10] Evaluation is more often defined as a discussion of the activities that has been carried out whether they have met the desired results or not, valuable or not, and to see the level of efficiency of its implementation.[11] Evaluation is also defined as a process that determines the conditions in which a goal has been achieved.[12] Evaluation is to render judgments about the value of whatever is being evaluated.[13]

According from the aspect of the evaluation function in education, evaluation is divided into

---

[10] Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment,* 41.

[11] Wahyuni and Ibrahim, 3.

[12] CD. Dirman and Cicih Juarsih. *Penilaian dan Evaluasi*. (Jakarta: Rineka Cipta, 2014), 32.

[13] Sefik Yasar and Asli Gundogan Cogenli, "Determining Validity and Reliability of Data Gathering Instruments used by Program Evaluation Studies in Turkey", Procedia Social and Behavioral Sciences (2014), 505.

two: first, evaluation can help the teacher in determining the degree of teaching goals to be achieved and second, evaluation can help the teacher for knowing the true state of the students. According to Arikunto, the functions of evaluation are as follows:

a. Evaluation has a selective function, namely to select students.
b. Evaluation has a diagnostic function that is to find out student weaknesses.
c. Evaluation has a placement function which is to find out where the student is placed according to his group
d. Evaluation has a function for measuring success

Then the purposes of evaluation are:

a. Evaluation is used to assess goal attainment
b. Evaluation is used to measure various types of aspects of learning that are varied.
c. Evaluation is used as a means (means) to find out what students already know
d. Evaluation is used to motivatie student learning.
e. Evaluation is used to provide information for learning purposes and errors
f. Evaluation is used to make the evaluation results as the basis for curriculum changes.

2. Assessment

Assessment is important things in education. Assessment is the process of collecting, synthesizing, and interpreting information to aid in

decision making.[14] Assessment is also defined as process of gathering information from the students about what they know and what they can do.[15] Then, after the teacher gets the information, the information will be used for making decision. Assessment must have two requirements, which are, assessment can measure the competence and must has beneficial effects toward teaching and learning process. The purposes of assessment are:

a. Assessment can make the teacher knowing the position of the students that is compared to other students.

[14] Russell and Airasian, *Classroom Assessment: Concept and Application*, 3.

[15] Wahyuni and Ibrahim, 2.

b. Assessment can separate between students who enter certain categories and who do does not belong to certain categories.

c. Assessment can give the teacher a description of the extent to which learners have mastered competence.

d. Assessment can be used to evaluate student learning outcomes in order to help students understand themselves and make decisions about the next step, both in the selection of personality development programs or in majors.

e. Assessment is used to indicate learning difficulties experienced by students and possible achievements that can be developed.

f. Assessment is used to obtain information that can predict how students perform at the next level of education.

g. Assessment is used to determine the level of efficiency of learning methods and other components used in a certain period of time.

Assessment is conducted continuously during the teaching and learning process. So, the teacher can conduct the process of assessment every time base on the method that teacher use. The procedure of conducting assessment according to Brown is divided into two, which are formative assessment and summative assessment.[16] Formative assessment is a process of unplanned assessment in the form of comments or responses. Then, summative

[16] Wahyuni and Ibrahim, 10.

assessment is an assessment process that is used to measure students' knowledge and abilities. These assessments are planned, systematic, and made for recognition of student achievement. The assessment is contained of measurement. Measurement is process of quantifying or assigning a number to a performance or trait.[17] Measurement will give the answer the question about how much. Therefore, the measurements will present quantitative results in the form of numbers or score. Measurement can be arranged in the form of test or non-test.

3. Test

Test is part of the measurement. A test, in simple terms, is a method of measuring a person's

[17] Russell and Airasian, *Classroom Assessment: Concept and Application*, 11.

ability, knowledge, or performance in a given domain.[18] Test is a tool or procedure that can be used to know or measure something in some conditions, with the rules.[19] Test is also defined as a way to organize evaluation in the form of tasks that must be done by the participant of the test.[20]

There are three types of the test:

a. Test Based on How to Do Test

Test based on how to do it are categorized into three which are: written test, spoken test, and action test.[21] Written test is the test that requires answers in the form of

[18] Brown. *Language Assessment Principles and Classroom Practices*, 3.

[19] Suharsimi Arikunto. *Dasar-dasar Evaluasi Pendidikan* (Jakarta: Bumi Aksara, 2013), 66.

[20] Wahyuni and Ibrahim, 11.

[21] Ibid, 11.

multiple choice or essay.[22] Spoken test is the test that requires answers orally that is conducted through face to face with the examiners.[23] Action test is the test which requires answers in the form of performance or action.[24] There are three kinds of action test which are: paper test and pencil test which is used to measure students' ability to display works; identification test which is used to measure students' ability to identify something; simulation test which is used to measure the mastery of students' skills with the help of artificial equipment or as if using a particular tool; and work sample test which is

[22] Ibid, 11.
[23] Ibid, 11.
[24] Ibid, 11.

used to measure the mastery of students' skills by using real equipment.[25]

b. Test Based on How to Answer Test

According how to answer, test is categorized into objective test and non-objective test. Objective test involves true-false, match test, and multiple-choice test.

1) True False Test

True-false test is the test that test items are typically written as statements, and students must decide whether the statements are true or false.[26] The advantages of true-false test are:

---

[25] Ibid, 11.

[26] Brown, *Testing in Language Programs : A Comprehensive Guide to English Language Assessment*, 47.

a) Many items can be administered in a relatively short time.

b) Moderately easy to write and easily scored.

Then, the weakness of true false testare:

a) limited primarily to testing knowledge of information.

b) Easy to guess correctly on many items, even if the material has not been mastered.

There are some suggestions for constructing true-false test as follows:

a) Test maker must include only one central, significant idea in each statement.

b) Test maker must word the statement so precisely that can unequivocally be judge true or false.

c) Test maker must keep the statement short, and use simple language structure.

d) Test maker must use negative statement sparingly, and avoid double negative.

e) Test maker must avoid extraneous clues to the answer.

f) The statements of opinion should be attributed to some sources.

2) Matching Test

Match test is the test that test items present the students with two columns of information; the students must then find and identify matches between two sets of information the sake of discussion, information given in left-hand column will be labeled matching item premise and that shown in the right-hand column will be labeled options.[27] In matching test, student

[27] Ibid, 50.

must match the correct option to the premise. The advantages of this test:

a) Items can be written quickly
b) A broad range of content can be assessed
c) Scoring can be done efficiently.

And the disadvantages of this test:

a) in aspect of cognitive skills, higher order cognitive skills are difficult to assess.
b) Focus is mainly on lower-level outcomes.
c) Homogeneous topics are required.

There are some suggestions for constructing matching test:

a) Test maker must include only homogeneous material in each matching item.
b) Test maker must keep the lists of items short and place the brief responses on the right.
c) Test maker should use a larger, or smaller, number of responses than premises, and permit the responses to be used more than once.
d) Test maker must specify in the directions the basis for matching, and indicate that

each response may be used once, more than once, or not at all.

3) Multiple-choice Test

Multiple-choice test is the test that test items are made up an item stem, or the main part of at the top, a correct answer, which is obviously the choice (usually, a,b,c,) that will be counted correct, and the distractors, which are those choices that counted as incorrect.[28] The most important requirement of a multiple-choice item is that the correct answer must be

[28] Ibid, 48.

genuinely correct.[29] The test maker also must give the key answer just one correct answer. Multiple-choice test is usually used in many kinds of assessment such as midterm examination, final examination, try-out test, etc.

The advantages of multiple-choice test are:

a) The test can be used to assess a broad range of content in a brief period.

[29] J. Charles Alderson, Caroline Clapam, and Diane Wall. *Language Test Construction and Evaluation*, (Trumpington: Cambridge University Press, 1995), 47.

b) Skillfully written items can be measure higher order cognitive skills.

c) The results can be scored quickly.

The disadvantages of this test are:

a) The arrangement of the test is difficult and time-consuming

b) The test is possible to assess higher-order cognitive skills, but most items assess only knowledge.

c) Some correct answers can be guesses.

d) Cheating may be facilitated in the test.[30]

There are some suggestions for constructing multiple-choice test as follow:

a) Test maker must design each item to measure an important learning outcome.
b) Test maker must present a single clearly formulated problem in the stem of the item.
c) Test maker must state the stem of the item in simple and clear language.

[30] Ibid, 78.

d) Test maker put as much of the wording as possible in the stem of the item.

e) Test maker must state the stem of the item in positive form, wherever possible.

f) Test maker must emphasize negative wording whenever it is used in the stem of an item.

g) Test maker must make certain that the intended answer is correct or clearly best.

h) Test maker must make all alternatives grammatically consistent with the stem of the item and paralel in form.

i) Test maker must avoid verbal clues that might enable students to select the correct answer or to eliminate an incorrect alternative.

j) Test maker must make the distractors plausible and attractive to the uniformed.

k) Test maker must vary the relative length of the correct answer to eliminate length as a clue.

l) Test maker must avoid using alternative “all of the above” and use “none of the above” with extreme caution.

m) Test maker must vary the position of the correct answer in a random manner.

n) Test maker must control the difficulty of the item either by varying the problem in the stem or by changing the alternatives.

o) Test maker must make certain each item is independent of the other item in the test.

p) Test maker must use an efficient item format.

Then, non-objective test is divided into three types, which involves completing test, short answer test, and essay test.

1) Completion Test

Completing test is the test that gives students essay questions which usually in the form of word or words. This test is also called fill-in the blank.[31] This test presents the student question to be answered and that question is in the form of an incomplete sentence, a picture, or a diagram that requires labeling. The advantages of this test are:

a) The opportunities for students to guess the answers of the test can be reduced.

[31] Russell and Airasian, *Classroom Assessment: Concept and Application*, 148.

b) The teacher can arrange question items easily.

c) Broad range of knowledge can be assessed.

The disadvantages of completion test are:

a) Scoring takes a long time

b) Not useful for complex or extended outcomes.[32]

2) Short Answer Test

Short answer test is the test that gives the students questions in form of word, words, or short sentences. The advantages of short answers test are:

[32] Ibid, 154.

a) The test items can be administered in a brief amount of time.

b) The test is relatively efficient to score.

c) The arrangement of the test items is moderately easy.

The disadvantages of short answer tests are:

a) Difficult to identify defensible criteria for correct answers.

b) Limited to questions that can be answered or completed in a few words.

There are some suggestions for conducting short-answer test:

a) Test maker must state the item so that only a single, brief answer is possible

b) Test maker must start with direct question, and switch to an incomplete statement only when greater conciseness is possible by doing so.

c) The words to be supplied should relate to the main point of the statement.

d) Test maker must place the blanks at the end of the statement.

e) Test maker must avoid extraneous clues to the answer.

f) For numerical answers, test maker must indicate the degree of precision expected and the units in which, they are to be expected.

3) Essay Test

Essay test is the test that the answer of the question requires students to remember and organize ideas or things they have learned by expressing them or expressing them in essay written form.[33] The characteristic of essay test is the freedom of response it provides.[34] The advantages of essay test are:

[33] Wahyuni and Ibrahim, 12.

[34] Norman E. Gronlund. *Constructing Achievement Test*. (New Jersey: Prentice Hall Inc., 1977), 75.

a) The test can be used to measure higher order cognitive skills.
b) The questions can be arranged easily.
c) The respondent is difficult to get the correct answer by guessing.

Then the disadvantages of essay test are:

a) Time-consuming to administer and score.
b) Difficult to identify reliable criteria for scoring.
c) Only a limited range of content can be sampled during any one testing period.

c. Test Based on Bloom's Taxonomy

Based on Bloom, test is divided into cognitive test, affective test, and psych motor test.[35] Cognitive is all activity that related to brain activity such as memorizing, interpreting, applying, problem solving, reasoning, analyzing, and thinking critically.[36] In cognitive there are six thinking process which are knowledge, comprehension, application, analysis, synthesis, and evaluation.[37] Knowledge is ability of the students to recall the information. Comprehension is the ability of the students to understand the explanation of the teacher. Application shows that how the students can applicant their knowledge in daily

[35] Russell and Airasian, *Classroom Assessment: Concept and Application*, 68.

[36] Ibid, 68.

[37] Ibid, 68.

activity. Analysis is the student ability to analyze the problems in details. Synthesis is the ability of students to combine elements of a thing into a comprehensive form that is structured or forms a new pattern. Evaluation is the ability of students to consider a situation, value, or idea.

Affective is all things related to attitudes, feelings, interests, preferences, emotions, and values.[38] The types of affective learning outcomes appear to students in various behaviors such as attention to learn, discipline, motivation to learn, respect to the teachers and classmates, study habits, and social relations. There are several categories of affective

[38] Ibid, 71.

domains include: receiving, responding, valuing, organizing, and characterization by value. Receiving is the willingness of students to show activities or receive stimulus, control or external stimulation in the form of a situation. Responding is a reaction from students to stimuli that come from outside on the basis of their own desires. Valuing is the ability of students to assess after being motivated by the teacher. Organizing is the development of the ability to assess toward an organizational system, including the relationship between the ability to assess certain values and the stabilization of values that have been owned. Characterizing by value is the integration of all the value systems that

have been owned by students that affect personality patterns and behavior.

Psychomotor domain is anything related to brain, physical or limb movements. Psychomotor learning outcomes are in the form of certain motion skills acquired after learning. Then, these skills are always associated with the movement of skills in accordance with the subjects of being taught. The assessment which is used in psychomotor is an action test. There are types of psychomotor test include: imitation, manipulation, precision, articulation, and naturalization. Imitation is the process of observing and determining the behavior patterns of others then imitating it.

Manipulation is an activity that requires students to be able to perform certain performance by following instructions and practice. Precision is an activity that requires students to be able to refine performance to be more precise. Naturalization is a demand for students to have a high level of performance, natural performance, and without thinking.

Based on the explanation above, there many kinds of the test. All of the test will show the quality that can show good result or bad result, if the test has been analyzed. Arikunto stated that a good test must have: validity, reliability, objectivity, practicability, and economist.[39] To analyze that the test is good or

[39] Arikunto, *Dasar-dasar Evaluasi Pendidikan*, 72.

not, the teachers need to analyze the quality of the test by conducting item analysis.

4. Criteria of the Good Test

Tests or test questions are measuring instruments that have multiple functions, which is used to measure learning effectiveness and measure the effectiveness of teachers in teaching. A test can be a good measurement tool and provide accurate information when the question as part of the testing construction is maintained quality. A good test as a measurement tool is a test that has criteria in the form of validity, reliability, objectivity, practicability, and economist.[40] According to Brown, a test can be said to be good if it has

[40] Ibid, 72.

validity, reliability, authenticity, practicability, and washback.[41]

The test can be said as good in validity, if it can show valid result in measuring result of the test or it has high validity. Then, the test can be said reliable, if the result of the test can be considered consistent or stable or it has high reliability. The next is practically, it means that the test has good procedure, efficient, and cheap. Therefore, the test has been structured before conducted. Then, authenticity of the test refers to the contents of the test. The contents of the test must be authentic or have natural language, relevant topic, and tasks represent. Washback refers to the effect of the test have on instructions in terms how of the students

[41] Brown, *Language Assessment Principles and Classroom Practices,* 19.

prepare for the test.[42] Washback is also defined as the concept as consequential perspectives related to score meaning and the intended and unintended consequences of assessment utilization.[43]

While, the quality of the test based on criteria of the good test can be analyzed by examining the test with certain questions, conducting the item analysis, by finding the reliability of the test, or conducting validity analysis.[44] This study uses test item analysis to determine the quality of the test. The test item analysis is conducted by calculating the aspect of validity degree, reliability degree, level of difficulty degree, discrimination index

[42] Ibid, 28.

[43] Sibel Toksoz and Ayse Ertunc, "Item Analysis of Multiple-Choice Exam", Advances in Language and Literary Studies, 8, (December 2017), 142.

[44] Arikunto, *Dasar-dasar Evaluasi Pendidikan*, 220.

degree, and distribution pattern answer or item distractor degree. The degree of each aspect can be determined by seeing the table below:

a. Validity

Table 1.1 Validity Degree[45]	
The Result of Validity	**Criteria**
r is between 0,8 until 1,0	Very high
r is between 0,6 until 0,8	High
r is between 0,4 until 0,6	Medium
r is between 0,2 until 0,4	Low
r is between 0,0 until 0,2	Very low

b. Reliability

Table 1.2 Reliability Degree[46]	
The Result of Reliability	**Criteria**
r is between 0,8 until 1,0	Very high
r is between 0,6 until 0,8	High
r is between 0,4 until 0,6	Medium
r is between 0,2 until 0,4	Low
r is between -1,0 until 0,2	Very low

c. Item Difficulty (Item Facility)

[45] Ibid, 89.

[46] Anita Noveria, "Item Analysis on the Validity and the Reliability of the English Summative Test for the first-year students at MA Madani Alauddin Pao-pao", *ISERD International Conference*, (March, 2018), 24.

Table 1.3 Item Difficulty Degree[47]	
The Result of Item Difficulty (IF)	**Criteria**
Item Facility (IF) is between 0.71-1.00	Easy
Item Facility (IF) is between 0.31-0.70	Medium
Item Facility (IF) is between 0.00-0.30	Difficult

d. Item Discrimination

Table 1.4 Item Discrimination Degree[48]	
The Result of Item Discrimination (ID)	**Criteria**
Item discrimination (ID) shows less than 0	Very bad (must be revised).
Item discrimination (ID) shows between 0 until 0.19.	Bad
Item discrimination (ID) shows between 0.20-0.29.	Medium
Item discrimination (ID) shows between 0.30-0.39.	Good
If item Discrimination (ID) shows more than 0.40.	Very good

e. Item Distractor

Table 1.5 Item Distractor Degree[49]

[47] Arikunto, *Dasar-dasar Evaluasi Pendidikan*, 225.
[48] Ibid, 136.

The Result of Item Distractor	Criteria
Item distractor is chosen by minimum 5% from the participant, distractor is good.	Good
Item distractor is chosen by less than 5% from the participant, distractor is not good	Bad

5. Test Item Analysis

Item analysis is the systematic evaluation of the effectiveness of the individual items on a test.[50] Item analysis is also defined as analysis which is conducted toward test items to find level of difficulty, item discrimination, and item distractor.[51] The test item analysis aims to reveal

---

[49] Hamzah B. Uno and Satria Koni. *Assessment Pembelajaran.* (Jakarta: PT Bumi Aksara, 2014), 158

[50] Brown, *Testing in Language Programs : A Comprehensive Guide to English Language Assessment*, 41.

[51] Uno and Koni, *Assessment Pembelajaran*, 156.

description, quality of test items, and things that related to development, arrangement, or application of the test that has appropriate and need to be maintained.[52] Item analysis is also used to identify either the question is a good, less good or not a good question. A not good question should be revised or discarded. So, the test is actually made up of the good quality questions.

According to Arikunto, the purposes of item analysis are: identifying bad test items, gathering information that can be used to revise questions test, and giving description about the test arrangement.[53] While, according to Tuckman, the purposes of item analysis are: making each question items consistent toward all test items and

[52] Wahyuni and Ibrahim, 128.
[53] Ibid, 128.

measuring the effectiveness of the test as instruments tools.[54] Furthermore, Silverius reveals the functions of item analysis as follows: first, item analysis is used to determine the accuracy of test items whether they are appropriate toward the wishes of test makers; second, giving feedback to the students according to the test results; third, getting feedback from the students related learning difficulties; fourth, improving curriculum and test items; and fifth, improving skills to compile the test.[55]

The benefits of item analysis are: making balance the proportion of items that easy, medium, and difficult in the test; extending and shortening the test in order to improve validity and reliability

[54] Ibid, 128.
[55] Ibid, 129.

of the test; simplifying the basic concept and item analysis technique; and supporting the test user to evaluate the test that has been published.[56]

In addition**,** an analysis of test items will obtain the important information, which basically would be useful feedback to make improvements, enhancements, and refinements to those items that have been issued in the achievement test, so in future tests of learning outcomes are arranged or designed by the evaluator who can evaluates learning outcomes that have good quality.

The test item analysis is conducted by calculating the aspect of validity, reliability, level of difficulty, discrimination index, and distribution

[56] Uno and Koni, *Assessment Pembelajaran*, 158.

pattern answer or item distractor. The explanation are as follows:

a. Validity

Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure.[57] Validity is also defined as the extent to which an instrument measured what it claimed to measure.[58] Validity suggests truthfulness and refers to the match between a construct, or the way a researcher conceptualizes the idea in a conceptual definition, and a measure. It refers to how well

[57] Grant Hening. *A Guide to Language Testing: Developtment, Evaluation, and Research.* (Foreign Language and Research Press, 2001), 89.

[58] Donald Ary, Lucy Cheser Jacobs, and Christine K. Sorensen. *Introduction to Research in Education.* (Belmont: Wadsworth, 2010), 225.

an idea about reality "fits" with actual reality. The test can be said as valid if it measures what it purposes to measure. The concept of validity, as used in testing can be clarified further by noting the following general points:

1) Validity refers to the interpretation of test results (not to the test itself)
2) Validity is inferred from available evidence (not measured)
3) Validity is specific to a particular use (selection, placement, evaluation of learning, and so forth).
4) Validity is expressed by degree (for example high, moderate, and low).[59]

[59] Norman E. Gronlund. *Constructing Achievement Test*. (New Jersey: Prentice Hall Inc., 1977), 131.

Test validity is defined as the degree to which a test measures what it claims, or purports, to be measuring.[60] Validity is divided into four which are: logical validity and empirical validity. Logical validity is used to measure the logical result. There are two kinds of logical validity that can be reached by an instrument which are:

a. Content Validity

Content validity is the accuracy of a measuring tool which is analyzed based on the content of the measuring instrument.[61] A test is said to have content validity if its contents constitutes a representative sample of the language skills, structure,

[60] Brown, *Testing in Language Programs*, 231.
[61] Wahyuni and Ibrahim, 86.

and etc. with which it is meant to be concerned.[62] The measurement tools can be said valid if the content of the measurement tools representative toward teaching material which is given. The content validity of the test results can be seen in the learning achievement test whether it is in accordance with the material that has been taught. Teacher can build a test that has high content validity by identifying the subject-matter topics and behavioral outcomes to be measured, building table of specification, which specifies the sample of item to be used,

[62] Arthur Hughes. *Testing for Language Teachers*. (Cambridge: Cambridge University Press, 2003), 26.

and constructing a test that closely fits the table of specifications.[63]

b. Construct Validity

Construct validity refers to the suitability between the results of the measuring instrument and the ability that want to be measured.[64] The measuring tool can be said to have validity construct if the score of a measuring instrument is appropriate with the ability that want to be measured. The aim in determining construct validity is to identify all the factors that influence test performance and to determine the degree of influence of

[63] Gronlund, *Constructing Achievement Test*, 132.
[64] Wahyuni and Ibrahim, 87.

each. The process includes the following steps:

1) Identifying the constructs that might account for test performance
2) Formulating testable hypotheses from the theory surrounding each construct
3) Gathering data to test these hypotheses

Then, empirical validity or criterion validity is used for measuring the instrument tools based on experience. The concept of criterion validity involves demonstrating validity by showing that scores on a test being validated correlate highly with some other, well-respected

measure of the same construct. There are two kinds of criterion validity which are:

a. Concurrent Validity

Concurrent validity refers to how far the test has similarities with existing or standardized tests.[65] Concurrent validity is concerned with the use of test performance to estimate current performance on some criterion[66]. Concurrent validity can be reached when an indicator is associated with a preexisting indicator that is judged to be valid with a

[65] Uno and Koni, *Assessment Pembelajaran*, 152.
[66] Gronlund. *Constructing Achievement Test*, 133.

preexisting indicator that is judged to be valid.[67]

b. Predictive Validity

Predictive validity refers to how far the test can determine or predict certain criteria that are targeted[68]. An indicator will predict future events that are logically related to a construct. Predictive validity is concerned with the use of test performance to predict future performance on some other valued measure called a criterion.[69]

[67] W. Lawrence Neuman. *Basic of Social Research: Qualitative and Quantitative Approach*. (Boston: Pearson Education ,2007), 118.

[68]Uno and Koni, *Assessment Pembelajaran*, 153.

[69] Gronlund. *Constructing Achievement Test*, 133.

The key element in both types of criterion validity is the degree of relationship between two sets of measures (the test score and the criterion to be predicted or estimated).[70]

Technique which can be used for analyzing the validity: product moment correlation (Pearson Method) and rank method of correlation (Spearman Method), point-biserial correlation.

Product moment correlation is technique of analysis correlation data between two variables.[71] Rank method of correlation is technique of analysis item validity using

[70] Ibid, 134.

[71] Retno Widyaningrum. *Statistika.* (Depok: Pustaka Felicha, 2015), 105.

relative ranking score of each score in two groups of participants.[72] Then, point-biserial correlation is technique of analysis item validity using nominal scale data, right or wrong answer test, or continuous scale. This study will use Pearson product moment correlation to analyze the data. The formula is:

$$(r_{xy} = \frac{\sum(X - M_x) - (Y - M_y)}{NS_xS_y})$$

r_{xy}= Pearson correlation coefficient

X = each student's score on test X

M_x= mean on test X

S_x= standard deviation on test X

[72] M. Ngalim Purwanto. *Prinsip-prinsip dan Teknik Evaluasi Pembelajaran*. (Jakarta: Remaja Rosdakarya, 2013), 146.

Y= each student's score on test Y

M_y= mean on test Y

S_y= standard deviation on test Y

N = number of the students

There are some tips to make the tests more valid:

1. Test maker must write explicit specifications for the test which take account of all that is known about the constructs that are to be measured.
2. Wherever feasible, test maker must use direct testing

3. Test maker must make sure the scoring of responses relates directly to what is being tested.
4. Test maker must do everything possible to make the test reliable.

b. Reliability

Reliability is essentially a synonym for consistency and replicability over time, over instruments and over groups of respondents.[73] According to Nana Sudjana, reliability refers to the accuracy or constancy of a measuring instrument to assess what is judged.[74]

---

[73] Louis Cohen, Lawrence Manion, and Keith Morrison. *Research Methods in Education.* (London: Taylor & Francis e-Library, 2005), 134.

[74] Noorrachma Chandra Novianti, "Test Item Analysis of the Final Examination on Economics Subject in Grade XII IPS SMA Negeri 1 Wonosari Academic Year 2014/2015", Thesis (Yogyakarta: Yogyakarta State University, 2015), 13.

Reliability can be defined as dependability or consistency.[75] It means that the numerical results produced by an indicator do not vary because of characteristics of the measurement processor measurement instrument itself. The test can be said reliable if the test can be trusted, consistent, stabile, and productive.[76] Two aspects of reliability are stability and equivalence.[77] There are two kinds of reliability, which are; internal reliability and external reliability.[78] Internal reliability is test of reliability that comparing test parts with

[75] Brown. *Language Assessment Principles and Classroom Practices*, 20.

[76] Purwanto, *Prinsip-prinsip dan Teknik Evaluasi Pembelajaran*, 139.

[77] C.R. Kothari. *Research Methodology Methods and Techniques.* (New Delhi: New Age Publishers, 2004), 75.

[78] Wahyuni and Ibrahim, *Asesmen Pembelajaran Bahasa*, 105.

each other in the test itself.[79] Then, external reliability is conducted by comparing test score with others test score/ different test.[80] Reliability test can be conducted by some methods. Internal reliability can be conducted by using test-retest method and equivalent-forms method.

1) Test-retest Method

Test-retest method is used to conduct reliability test of instrument tools by testing instrument tools twice or more, and the result will be correlated.[81] The purpose of this reliability test is to know the stability coefficient of the instrument

[79] Ibid, 105.
[80] Ibid, 105.
[81] Ibid, 106.

tools. There are steps to conduct this reliability test:

a) Tester arranges the instrument tools that will be measured that reliability.
b) Tester tests the instrument tools that had been arranged.
c) Tester calculates the result of the first test from instrument tools.
d) Tester tests the instrument tools that had been arranged.
e) Tester calculates the result of the repetition test from instrument tools
f) Tester calculates the reliability of the instrument tools by correlating the result of the first test and the repetition test from instrument tools

using Pearson formula or product-moment formula. The formula is:

$$(r_{xy} = \frac{\sum(X - M_x) - (Y - M_y)}{NS_xS_y})$$

r_{xy}= Pearson correlation coefficient

X = each student's score on test X

M_x= mean on test X

S_x= standard deviation on test X

Y= each student's score on test Y

M_y= mean on test Y

S_y= standard deviation on test Y

N = number of the students

2) Equivalent-forms Method

Equivalent-forms method is conducted by arranging two measuring instruments that have similarities, parallels, or equivalents.[82] After both measuring instruments are tested, the results of the measuring instruments test are correlated. This method is used to know the stability coefficient of the instrument tools by assuming that the system which is measured will not change with only two measuring instruments used**.** There steps for conducting this method:

a) Tester arranges two instrument tools which both equivalent

[82] Wahyuni and Ibrahim, *AsesmenPembelajaran Bahasa*, 108.

b) Tester test two instrument tools in the same time

c) Tester gives the score results of the tested measuring instrument, arranged by separating between measuring instrument A and measuring instrument B.

d) Tester calculates the stability coefficient of the two measuring instruments by determining correlations with the formula.

Then, external reliability can be conducted using split-half method, and homogeneity method.

1) Split-half Method

This method is conducted by splitting the instrument tools, for example, the test is divided into two and the score of two parts of the test is correlated using formula.[83] There are two ways to do this method, which are splitting between odd scores and even scores or by splitting the top and bottom numbers.

The opinion that underlies the use of this method is that a test is arranged in a systematic pattern, so that if it is divided according to even odd or bottom up, it will not change the score position of each student. There are steps for conducting this test:

[83] Ibid, 108.

a) Tester compile a test with even number of numbers

b) Tester test the test

c) Tester calculates the test score of each sample, by grouping it into even odd or bottom-up scores.

d) Tester determines reliability by correlating the two scores with the product-moment formula.

e) Tester determines reliability all of the test Spearman-Brown formula, Flanagan, or Rulon.

There are formulas which can be used for this method: Spearman-Brown (split-half), Flanagan, or Rulon.

This research uses the Spearman-Brown for calculating reliability analysis. The formula is:

$$(r_{crt} = \frac{2.rb}{1+rb})^{84}$$

r_{crt} = coefficient reliability of total test

rb = correlation between the scores of each parts of the test

2) Homogeneity Method

This method is used to analyze reliability test that cannot be tested by spilt-half method.[85] There are the formulas:

a) Kuder Richardson (K-R 20)

[84] H.M. Sukardi, *Evaluasi Pendidikan: Prinsip dan Operasionalnya.* (Jakarta: BumiAksara, 2011), 48.

[85] Wahyuni and Ibrahim. *Asesmen Pembelajaran Bahasa*, 108.

This formula is invented by Kuder Richardson. The use of this formula: first, tester makes item analysis tables without having to group them into odd numbers and even numbers; second, tester calculates the proportion that answers correctly and the proportion that answers incorrectly on each item in the item analysis table; third, tester is multiplying the proportions that answer correctly and the proportion that answers incorrectly; fourth, tester calculates variance (standard deviation squared) from the total score; fifth, tester calculates the

reliability of the test with the formula K-R 20.[86] K-R 20 formula:

$$r_{11} = \left\{\frac{k}{k-1}\right\}\left\{\frac{SDt^2 - \sum pq}{SDt^2}\right\}$$

r_{11} = reliability of the test

k= the total number of items

p= the proportion of subjects who answered correctly in each item

q= the proportion of subjects who answered incorrectly in each item

$\sum pq$= the total number of p and q of each item that has been multiplied

b) Kuder Richardson (K-R 21)

[86] Ibid, 116.

This formula can be conducted by knowing total scores and variants from that scores, number item totals and mean.[87] K-R 21 formula:

$$r_{11} = \left\{\frac{k}{k-1}\right\}\left\{\frac{M(k-M)}{kSDt^2}\right\}$$

r_{11} = reliability of the test

k= the total number of items

M= mean of the scores

c) Hoyt

This formula is used to measure the reliability of the test which the scoring method is 1 and 0.[88] The steps for conducting this formula: first,

[87] Ibid, 118.
[88] Ibid, 119.

tester determines the square numbers of respondents; second, tester calculates the sum of squares from the items; third, tester calculates for the sum of the remaining squares; tester determines residual variance using table F; tester inserts the data into formula R_{11} Hoyt.

d) Alpha

Alpha formula is used to determine reliability of the test which use Likert scale or essay test.[89]

There are some tips to make the tests more reliable:

[89] Ibid, 121.

1. Test maker must take enough samples of behavior.
2. Test maker must exclude items which do not discriminate well between weaker and stronger students.
3. Test maker must not allow candidates too much freedom.
4. Test maker must write unambiguous items.
5. Test maker must provide clear and explicit instructions.
6. Test maker must ensure that tests are well laid out and perfectly legible.
7. Test maker must make candidates familiar with format and testing techniques.

8. Test maker must provide uniform and non-distracting conditions of administration.
9. Test maker must use items that permit scoring which is as objective as possible.
10. Test maker must provide a detailed scoring key.
11. Test maker must train scorers
12. Test maker must agree acceptable responses and appropriate scores at outset of scoring.
13. Test maker must identify candidates by number, not name.
14. Test maker must employ multiple and independent scoring.

c. Item Difficulty

A good test is the test that has question not too easy or not too difficult.[90] Very easy questions do not give students stimulus to enhance the effort to do it. In contrast, very difficult questions will make students give up and not eager to do it. From that matter, test makers need to conduct level of difficulty analysis for revealing the level of difficulties of the test. The level of difficulty of the test is analyzed using item facility formula. Item difficulty also known item facility is a statistic used to examine the percentage of students who correctly answer given item.[91] Item facility also show how difficult or easy test

[90] Ibid, 129.

[91] Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*, 66.

items that have been held.[92] To calculate item facility, add up the number of students who correctly answers a particular item and divide that sum by the total number of students who takes the tests.[93] The formula that can be used to calculate item facility is:

$$(\text{IF}= \frac{N_{correct}}{N_{total}})^{94}$$

$N_{correct}$: number of students answering correctly

N_{total} : total number of students taking the test

The interpretation of the formula is: If the result of item facility analysis (IF) shows

[92] Wahyuni and Ibrahim. *Asesmen Pembelajaran Bahasa*, 129.
[93] Ibid, 66.
[94] Ibid, 66.

the value more than 0.96, item is easy. Then, if the result of item facility analysis (IF) shows the value between 0.28-0.95, item is medium. Then, if the result of item facility analysis (IF) indicates the value less than 0.27, item is difficult.[95] Therefore, if more students answer the test correctly, the test can be said to be easier.

d. Item Discrimination

Item discrimination is a statistics which indicates the degree to which item separates the students who performes well from those who did poorly on the test as a whole. Item discrimination also shows the difference of the skills between the student who has high ability

[95] Ibid, 67.

and the student who has low ability.[96] Item discrimination can be calculated using the formula as follow:

$$(\text{ID} = IF_{upper} + IF_{lower})^{97}$$

ID = item discrimination for an individual item

IF_{upper} = item facility for the upper group on the whole test

IF_{lower} = item facility for the lower group on the whole test

The interpretation of the formula is: if the finding shows discrimination index (ID) less than 0, item is very bad (must be revised).

[96] Uno and Koni, *Assess,ment Pembelajaran*, 157.
[97] Wahyuni and Ibrahim, *Asesmen Pembelajaran Bahasa*, 69.

Then, if the finding shows discrimination index (ID) between 0 until 0.19, item is bad. Then, if the result of the research shows item discrimination index (ID) between 0.20-0.29, item is medium. Then, if the result of the test shows item discrimination index (ID) between 0.30-0.39, item is good. And if the finding shows item discrimination index (ID) more than 0.40, item is very good.[98]

e. Item Distractor

The distribution pattern of answer can be obtained by counting the number of test participants who choose the answer of a, b, c, d, or do not select anything. From the distribution pattern of answer can be obtained

[98] Ibid, 136.

information about whether detractors are functioning properly or not. The distractors can function well if at least chosen by 5% of all learners who participate in test. The formula which is used to identify the quality of distractor:

$$IDs = 5\% \times n_{ts}$$

IDs = percentage

n_{ts} = total of the students who took the test.

6. Try-Out Test (Computer-Based Test)

a. Definition of Try-Out

Try-out is the test that is used to measure the student's ability in mastery of certain

material subjects.[99] Try-out is also defined as a mechanism that is used as an exercise for students before conducting the real exam.[100] Try-out can be used as input for the teacher to find the right and fast strategy in providing understanding to students on certain indicators.[101] Try-out is also used as an effort to measure the ability of students to be able to provide an overview of the competencies that have been achieved or mastered by students, and provide experience to students in carrying out tests as will be done on the National

[99] Yulia Elfiza, Rusman, and M. Nasir," Hubungan antara Hasil Uji Kognitif Try Out Ujian Nasional (UN) dengan Hasil Ujian Nasional (UN) Mata Pelajaran Kimia SMA Kota Banda Aceh Tahun Ajaran 2014/2015, "Jurnal Ilmiah Mahasiswa Pendidikan Kimia, 1, (2016), 36.

[100] Ibid, 36.

[101] Ibid, 37.

Examination.[102] In Indonesia, try out is conducted by schools before facing the National Examination.

b. Computer-Based Test

The first computers are introduced as a new technology which has the potential to generate types of learning environments, new settings for the design and administration of tests. Computer is often referred to as a tool that can enhance instruction.[103] In reality, computers are more like a toolbox than a single tool. From the development of the computer system, Computer-Based Test (CBT) is

[102] Arrynda Ratnaningsih, "Analysis Kualitas Soal-Soal Try Out Ujian Nasional Mata Pelajaran IPA SMP di Kabupaten Banjarnegara", Thesis (Semarang: Semarang Stated University, 2012), 9.

[103] Russell and Airasian, *Classroom Assessment: Concept and Application*, 337.

introduced as an assessment test tool. CBT is known as Computer-Based Assessment or e-exam. It is a method of administering tests in which the responses are electronically recorded, assessed, or both.[104] Compared to the well-known paper-and-pencil tests, computer-based test offers a number of advantages. Due to the interactive testing environments and the data-processing capacities offered by the computer, these potential advantages and added values cover a range of new possibilities that might extend from standardized and automatized administration and scoring procedures through interactive and media-

[104] A.T. Alabi, A. O. Issa, and R. A. Oyekunle, "The Use of Computer Based Testing Method for the Conduct of Examinations at the University of Ilorin*," International Journal of Learning & Development*,3(April-May, 2012), 69.

enriched new item types to the possibility of recording and exploiting behavioral data or the possibility of new test administration procedures such as adaptive testing. The computer offers the possibility to simulate the test environments, to record important interactions, and thus to administer tests for these interactive constructs in an efficient and scalable way.

The main value of CBT historically has been in the area of report generation. Some of the earliest computer systems are designed to automate the scoring and interpretation of instruments. With the advent of the personal computer, the development of computer-administered versions of paper and pencil tests

has been conducted which provides some advantages between paper and pencil, in terms of control of administration.[105]

The use of computer-based testing is increasing rapidly. This increase has been helped not only by the development of better interfaces, but also increase in the volume and accessibility to hardware or software. After the availability of internet access, CBT gives some offers in using the internet to support test users, assessment for development, test practice and familiarization, recruitment and selection post-hire applications, including online 360-degree feedback and development.

[105] Dave Bartram and Ronald K. Hambleton. *Computer-Based Testing and the Internet Issues and Advances*. (Chichester: John Wiley and Son Ltd., 2006), 13.

Computer-Based Test (CBT) will continue to grow with increasingly innovative offers.

C. Theoretical Framework

Evaluation demands teachers to make an assessment in the form of a good test. A good test must have validity, reliability, objectivity, practicability, and economist. The evaluator can make a good test by analyzing the quality of the test using a test item analysis. An item analysis will help evaluator to develop the test.

This study will useful in improving the quality of the test by the evaluator. This study will be conducted by testing the validity, reliability, discrimination index, level of difficulty, and item distractor.

An analysis of each item aims to obtain the important information, which basically would be an

useful feedback to make improvements, enhancements, and refinements toward items that have been issued in the try-out test, so in future tests of learning outcomes are arranged or designed by the evaluator who can evaluate learning outcomes that have good quality.

CHAPTER III

RESEARCH METHODOLOGY

In this chapter, the researcher discusses about research design, population and sample, instrument of data collection, technique of data collection, and technique of data analysis.

A. Research Design

Good ideas are very important in the research and the ideas must be backed up by good design.[106] Therefore, the research needs a research design. Research design is the arrangement of conditions for collection and analysis data in a manner that aims to

[106] Mark Balnaves and Peter Caputi, *Introduction to Quantitative Research Methods: An investigative Approach*, (London: Sage Publicaions, 2001), 27.

combine relevance to the research purpose with economy in procedure.[107]

This study is descriptive research which is included as the quantitative research. Descriptive research is the research that includes surveys and fact-finding enquiries of different kinds.[108] The major purpose of descriptive research is the description of the state of affairs as it exists at present.[109] In practice, this research intends to seek information and data that can be used to describe the quality of the test in MTs Negeri 2 Ponorogo. While the data obtained will be realized in the numerical form of figures and analyzed using IBM SPSS program version 23.

---

[107] C.R. Kothari, *Research Methodology Methods and Techniques.* (New Delhi: New Age Publishers, 2004), 31.

[108] Rafikasari Risqi Sakti, "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 (Ponorogo: Stated institute of Islamic Studies Ponorogo,2019), 35.

[109] Ibid, 35.

B. Instrument of Data Collection

Instrument is tool for assist which is chosen and used by the researcher in gathering data. Instrument of data collection can be shown as the table below:

Table 1.6 Instrument of Data Collection

Title of the Research	Variable	Indicator	No. Item of Instrument
Item Analysis of Try-out Test (Computer-Based Test) at MTs N 2 Ponorogo	Validity	• If r_{xy} is between 0,8 until 1,0, item has very high validity. • If r_{xy} is between 0,6 until 0,8, item has high validity. • If r_{xy} is between 0,4 until 0,6, item has medium	1-50

		validity. • If r_{xy} is between 0,2 until 0,4, item has low validity. • If r_{xy} is between 0,0 until 0,2, item has very low validity.[110]	
	Reliability	• If r_{crt} is between 0,8 until 1,0, item has very high reliability. • If r_{crt} is between 0,6 until 0,8, item has very high reliability. • If r_{crt} is between 0,4	1-50

[110] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 89.

		until 0,6, item has very high reliability. • If r_{crt} is between 0,2 until 0,4, item has very high reliability. • If r_{crt} is between -1,0 until 0,2, item has very low reliability.[111]	
	Item difficulty	• If Item Facility (IF) between 0,71-1,00, item is easy • If Item Facility (IF) between 0.31-0.70,	1-50

[111] Anita Noveria, "Item Analysis on the Validity and the Reliability of the English Summative Test for the first-year students at MA Madani Alauddin Pao-pao", 24.

		item is medium • If Item Facility (IF) between 0.00-0,30, item is difficult.[112]	
	Item discrimination	• If item discrimination (ID) less than 0, item is very bad (must be revised). • If item discrimination (ID) between 0 until 0.19, item is poor. • If item discrimination (ID) between 0.20-0.29, item is medium.	1-50

[112] Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 225.

		• If item discriminati on (ID) between 0.30-0.39, item is good. • If item Discriminati on (ID) more than 0.40, item is very good.[113]	
	Item distractor	• If each item distractor is chosen by minimum 5% from the participant, distractor is good. • If each item distractor is chosen by less than 5% from the participant,	1-50

[113] Ibid, 136.

		distractor is not good.[114]	

C. Technique of Data Collection

The data collection technique is a way to obtain data in accordance with the type of data required. In this research, the data collection technique used is the documentation. This technique is used to get the questions of try-out test in English subject of IX grade MTs N 2 Ponorogo with answer keys and the answers of all students in grade.

D. Technique of Data Analysis

The questions of try-out test are in the form of multiple-choice or objective analyzed using test item analysis. Before analyzed firstly did the scoring for each answer of learners. Scoring scale is 0-1, a score of

[114] Uno and Koni, *Assessment Pembelajaran*, 180.

0 for incorrect answers, while a score of 1 for the correct answer. The data are then analyzed include:

1. Validity

Test validity is defined as the degree to which a test measures what it claims, or purports, to be measuring.[115] Analysis of validity will show the instruments of the test valid or invalid. Test can be declared valid if it has significant correlation toward total score. Product moment correlation will be used for calculating validity analysis. The formula is:

$$(r_{xy} = \frac{\sum(X - M_x) - (Y - M_y)}{NS_xS_y})$$

r_{xy}= Pearson correlation coefficient

[115] Brown, *Testing in Language Programs*, 231.

X = each student's score on test X

M_x= mean on test X

S_x= standard deviation on test X

Y= each student's score on test Y

M_y= mean on test Y

S_y= standard deviation on test Y

N = number of the students

The calculation of validity test is using IBM SPSS program version 23.

2. Reliability

Reliability means dependability or consistency.[116] High and low reliability of the test

[116] Brown. *Language Assessment Principles and Classroom Practices,* 20.

can be determined by looking at the size of the coefficient of reliability of the test. Reliability of the test is calculated using Spearman-Brown (split-half) formula. The formula is:

$$(r_{crt} = \frac{2.rb}{1+rb})^{117}$$

r_{crt} = coefficient reliability of total test

rb = correlation between the scores of each parts of the test

To calculate reliability test, the researcher uses IBM SPSS program version 23.

3. Item Difficulty

The difficulty level is an opportunity to answer a question correctly on the certain level of

[117] Sukardi, *Evaluasi Pendidikan: Prinsip dan Operasionalnya*, 48.

capabilities that are usually expressed in the form of an index. The test item can be expressed as a good item if the item is not too difficult and not too easy, in other words the level of difficulty is medium or sufficient. Item difficulty of the test is calculated using item facility formula. The formula is:

$$(IF=\frac{N_{correct}}{N_{total}})$$[118]

$N_{correct}$: number of students answering correctly

N_{total} : total number of students taking the test

The researcher uses IBM SPSS program version 23 to calculate item difficulty.

[118] Ibid, 66.

4. Item Discrimination

Item discrimination is a statistic which indicates the degree to which n item separates the students who performed well from those who did poorly on the test as a whole. Item discrimination of the test will be calculated using Item discrimination formula. The formula is:

$$(\text{ID} = IF_{upper} + IF_{lower})^{119}$$

ID = item discrimination for an individual item

IF_{upper} = item facility for the upper group on the whole test

IF_{lower} = item facility for the lower group on the whole test

[119] Ibid, 69.

The researcher uses IBM SPSS to calculate item discrimination.

5. Item Distractor

Item distractor is the information about the distribution pattern of answer whether distractor answers are functioning properly or not. The distractors can be said as good, if 5% of the participants choose the distractors. The researcher will calculate item distractor using formula:

$$IDs = 5\% \times n_{ts}$$

IDs = percentage

n_{ts} = total of the students who took the test.

CHAPTER IV

RESULT OF THE STUDY

In this chapter, the researcher discusses about research location, data description, data analysis, and interpretation and discussion.

A. Research Location

Research location is place where the research is conducted. In this thesis, the research is conducted in MTS N 2 Ponorogo. Here are preview of the research location:

1. History of MTS N 2 Ponorogo

MTs N 2 Ponorogo is a formal educational institution that establishes rankings with Junior High School and is commonly referred to as Junior High School which is characterized by Islam which

is established or managed by the Ministry of Religion.

The Madrasah is established based on the Decree of the Minister of Religion of the Republic of Indonesia Number: 27 of 1980 dated May 31, 1980 concerning Relocation of Public Madrasah and Teachers of Religion of the State. The Madrasah is supported by an Operational Permit from the Office of the Ministry of Religion, Ponorogo Regency number: MTs / 2283/2010 on July 1, 2015.

Since November 2016 through the Decree of the Minister of Religion of the Republic of Indonesia Number 673 Year 2016 dated 17 November 2016, the name of the Madrasah

Tsanawiyah of the State of Ponorogo has changed to Madrasah Tsanawiyah Negeri 2 Ponorogo.

Since the establishment of MTs N 2 Ponorogo until now, it has experienced the change of leadership of great figures as follows:

a. H. Muslim, BA
b. Drs. Abdullah
c. H. Kustho, BA
d. Drs. Sumardi Al Basyari
e. Drs. H. Imam Asngari, SH, MPd.
f. Drs. H. Sutarto Kerim
g. Drs. MochHaris, M. Pd. I
h. Drs. Tarib, M.Pd.I

Under the leadership of the principals of the madrasah above, the Ponorogo State Islamic Primary School shows an increase in quality and

existences in religious character education. And the hope that with increasing age, the madrasah will be able to make the best contribution to the cause of Islam and the progress of science and technology based on Imtaq's stability.

2. The Geographical Location of MTS N 2 Ponorogo

MTs N 2 Ponorogo is located at Ki Ageng Mirah Street No. 20 Japan, Babadan, Ponorogo.

3. The Vision and Mission of MTS N 2 Ponorogo

The visions of MTs N 2 Ponorogo are building Indonesian Muslims with noble character, global outlook, smart, skilled, having the meaning, good science and technology, cared, cultured, and environmentally friendly.

The missions of MTs N 2 Ponorogo are:

1) Developing attitudes and behaviors that are Islamic in nature and cultural values of the nation in the real life.
2) Developing an international standard curriculum for MIPA, English and Arabic languages by adopting or adapting curriculum from developed countries.
3) Conducting teaching and learning process using multi resources and based on technology, information and communication (ICT).
4) Conducting the learning process actively, innovatively, creatively, effectively, cooperatively, communicatively, and inspiring students.
5) Growing the spirit of concern for the social environment, physical environment, and

instilling frugal life in environmental conservation efforts.

6) Applying a culture of clean living in order to prevent environmental pollution in everyday life.
7) Familiarize polite behavior in an effort to prevent environmental damage
8) Growing the spirit of competing in various competencies for all citizens of the madrasah.
9) Developing the potential and creativity of superior school citizens and able to compete both at regional, national and international levels.
10) Implementing Madrasah-Based School Management (MSBM) in a professional manner and leading to standardized education

quality management involving all members of the madrasa and other relevant institutions in the form of MoU.

11) Establish partnerships with superior schools / madrasah and universities as a companion for institutional development, human resources, curriculum and teaching and learning activities in the form of MoU.
12) Caring, cultured and environmentally friendly, being polite towards the environment by implementing caring in everyday life.

4. The Purposes of MTs N 2 Ponorogo

The purposes of MTs N 2 Ponorogo are:

1) Fulfill the needs of students who possess specific characteristics in terms of cognitive and effective development

2) Fulfill the human rights of students according to their educational needs
3) Fulfill the intellectual interests and future perspectives of students
4) Fulfill students' self-actualization needs
5) Considering the role of students as community assets and community needs for filling roles
6) Prepare students as future leaders
7) Producing higher quality output and outcomes for Ponorogo MTsN
8) Member opportunity for students who have the ability above average to complete the study program.

B. Data Description

Data that have been analyzed by the researcher are the answer sheets of English try-out test of the students. The data are result of the try out test. There are multiple choice test which contain of 50 questions. Each question has 4 choices answers which are A, B, C, and D.

C. Data Analysis

1. Validity

Formula that uses to show validity is product moment correlation proposed by Pearson. The formula is:

$$(r_{xy} = \frac{\sum(X - M_x) - (Y - M_y)}{NS_xS_y})$$

r_{xy}= Pearson correlation coefficient

X = each student's score on test X

M_x= mean on test X

S_x= standard deviation on test X

Y= each student's score on test Y

M_y= mean on test Y

S_y= standard deviation on test Y

N = number of the students

The validity is calculated using SPSS program. The objects in this research are 298. So, (r_{table}) is 0,113. The result of the test based on the degree of validity is as follows:

Table 1.7 The Result of Validity Test

No. Item	r_{xy}	r_{table}	Explanation
1	0,137	0,113	Valid (Very low)
2	0,296	0,113	Valid (Low)
3	0,184	0,113	Valid (Very low)

4	0,444	0,113	Valid (Medium)
5	0,395	0,113	Valid (Low)
6	0,382	0,113	Valid (Low)
7	0,331	0,113	Valid (Low)
8	0,501	0,113	Valid (Medium)
9	0,294	0,113	Valid (Low)
10	0,482	0,113	Valid (Medium)
11	0,328	0,113	Valid (Low)
12	-0,02	0,113	Invalid
13	0,459	0,113	Valid (Medium)
14	0,473	0,113	Valid (Medium)
15	0,336	0,113	Valid (Low)
16	0,096	0,113	Invalid
17	0,356	0,113	Valid (Low)
18	0,356	0,113	Valid (Low)
19	0,524	0,113	Valid (Medium)
20	0,126	0,113	Valid (Low)
21	-0,081	0,113	Invalid
22	0,589	0,113	Valid (Medium)
23	0,486	0,113	Valid (Medium)
24	0,165	0,113	Valid (Low)
25	0,335	0,113	Valid (Low)
26	0,437	0,113	Valid (Medium)
27	0,271	0,113	Valid (Low)
28	0,232	0,113	Valid (Low)
29	0,219	0,113	Valid (Low)
30	0,336	0,113	Valid (Low)
31	0,304	0,113	Valid (Low)
32	0,428	0,113	Valid (Medium)
33	0,501	0,113	Valid (Medium)
34	0,606	0,113	Valid (High)

35	0,388	0,113	Valid (Low)
36	0,529	0,113	Valid (Medium)
37	0,524	0,113	Valid (Medium)
38	0,393	0,113	Valid (Low)
39	0,224	0,113	Valid (Low)
40	0,081	0,113	Invalid
41	0,41	0,113	Valid (Medium)
42	0,443	0,113	Valid (Medium)
43	0,684	0,113	Valid (High)
44	0,569	0,113	Valid (Medium)
45	0,315	0,113	Valid (Low)
46	0,433	0,113	Valid (Medium)
47	0,591	0,113	Valid (Medium)
48	0,481	0,113	Valid (Medium)
49	0,316	0,113	Valid (Low)
50	0,308	0,113	Valid (Low)

From the table above, there are 4 invalid items, they are number: 12,16,21, and 40. Then, there are 2 numbers with very low validity, they are number: 1 and 3. Then there are 23 numbers with low validity, they are number: 2,5,6,7,9,11,15,17,18,20,24,25,27,28,29,30,31,35,38,39,45,49, and 50. Then, there are 19 numbers

with medium validity, they are number: 4,8,10,13,14,19,22,23,26,32,33,36,37,41,42,44,46,47, and 48. And then, there are 2 numbers with high validity, they are number: 34 and 43.

2. Reliability

Formula which is used in this research for showing reliability is Spearman-Brown (split-half). The formula is:

$$(r_{crt} = \frac{2.rb}{1+rb})^{120}$$

r_{crt} = coefficient reliability of total test

rb = correlation between the scores of each parts of the test

[120] H.M. Sukardi, *Evaluasi Pendidikan: Prinsip dan Operasionalnya*, 48.

The reliability is analyzed using IBM SPSS program. The result is:

Table 1.8 Case Processing Summary

		N	%
Cases	Valid	298	100.0
	Excluded[a]	0	.0
	Total	298	100.0

a. Listwise deletion based on all variables in the procedure.

The table above shows that there are 298 data that have been inputted. Then, all of the data input in this research are valid.

Table 1.9 Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
item_01	22.70	57.261	.102	.868
item_02	22.82	56.106	.247	.866
item_03	22.83	56.775	.132	.868
item_04	23.34	54.898	.394	.864
item_05	23.35	55.258	.344	.864
item_06	22.80	55.668	.338	.865
item_07	23.48	56.082	.289	.865
item_08	22.96	54.369	.452	.862

item_09	22.95	55.863	.236	.867
item_10	22.92	54.616	.434	.863
item_11	22.79	56.017	.283	.865
item_12	23.41	58.034	-.073	.871
item_13	22.93	54.729	.409	.863
item_14	23.13	54.393	.420	.863
item_15	22.79	56.001	.292	.865
item_16	23.36	57.296	.038	.870
item_17	23.48	55.981	.315	.865
item_18	22.85	55.657	.306	.865
item_19	22.89	54.415	.480	.862
item_20	23.44	57.150	.075	.869
item_21	23.44	58.375	-.132	.872
item_22	22.92	53.849	.548	.861
item_23	22.88	54.713	.440	.863
item_24	23.46	56.957	.117	.868
item_25	23.43	55.882	.287	.865
item_26	22.81	55.313	.394	.864
item_27	23.51	56.520	.231	.866
item_28	23.38	56.411	.177	.867
item_29	23.46	56.640	.171	.867
item_30	22.84	55.821	.287	.865
item_31	23.48	56.244	.261	.866
item_32	23.42	55.295	.383	.864
item_33	23.41	54.781	.459	.863
item_34	22.94	53.683	.565	.860
item_35	23.14	55.044	.330	.865
item_36	22.96	54.170	.483	.862
item_37	23.53	55.536	.496	.863
item_38	22.87	55.369	.343	.864
item_39	23.34	56.407	.167	.868
item_40	23.38	57.402	.024	.870

item_41	23.50	55.793	.373	.864
item_42	22.93	54.844	.392	.864
item_43	23.13	52.790	.646	.858
item_44	22.91	54.036	.527	.861
item_45	23.49	56.224	.274	.866
item_46	22.86	55.116	.386	.864
item_47	23.29	53.729	.548	.860
item_48	22.84	54.919	.438	.863
item_49	23.36	55.813	.263	.866
item_50	23.45	56.121	.262	.866

The table above shows the result of reliability analysis each item.

Table 1.10 Reliability Statistics

Cronbach's Alpha	N of Items
.867	50

From the table above, the result of $\propto$ scale is 0,867. Then, r_{table} is 0,113.

3. Item Difficulty

The formula which is used to calculate item difficulty is:

$$(IF=\frac{n_{correct}}{n_{total}})^{121}$$

$n_{correct}$: number of students answering correctly

n_{total} : total number of students taking the test

Item difficulty of the test is analyzed using SPSS program. If item facility (IF) is more than 0.96, item is easy. Then, if item facility (IF) is between 0.28-0.95, item is medium. Then, if item facility (IF) is less than 0.27, item is difficult.[122] The results are:

Table 1.11 Item Difficulty Analysis Results

No. Item	Item Difficulty	Interpretation
1	0,92	Easy
2	0,80	Easy
3	0,80	Easy
4	0,29	Difficult
5	0,28	Difficult

[121] Ibid, 66.

[122] Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*, 67.

6	0,82	Easy
7	0,15	Difficult
8	0,66	Medium
9	0,67	Medium
10	0,71	Easy
11	0,83	Easy
12	0,21	Difficult
13	0,69	Medium
14	0,49	Medium
15	0,84	Easy
16	0,26	Difficult
17	0,14	Difficult
18	0,78	Easy
19	0,73	Easy
20	0,18	Difficult
21	0,18	Difficult
22	0,70	Medium
23	0,74	Easy
24	0,16	Difficult
25	0,19	Difficult
26	0,82	Easy
27	0,12	Difficult
28	0,24	Difficult
29	0,17	Difficult
30	0,79	Easy
31	0,14	Difficult
32	0,20	Difficult
33	0,21	Difficult
34	0,69	Medium
35	0,48	Medium
36	0,67	Medium

37	0,09	Difficult
38	0,76	Easy
39	0,29	Difficult
40	0,24	Difficult
41	0,12	Difficult
42	0,69	Medium
43	0,49	Medium
44	0,71	Easy
45	0,13	Difficult
46	0,76	Easy
47	0,33	Medium
48	0,79	Easy
49	0,27	Difficult
50	0,17	Difficult

From the table above, there are 16 easy test items, they are number: 1,2,3,6,10,11,15,18,19,23, 26,30,38,44,46, and 48. Then, there are 11 medium test items, they are number: 8,9,13,14,22,34,35,36, 42,43, and 47. Then, there are 23 difficult test items, they are number: 4,5,7,12,16,17,20,21,24,25, 27,28,29,31,32,33,37,39,40,41,45,49, and 50.

4. Item Discrimination

The formula which is used to calculate item discrimination is:

$$(\text{ID} = IF_{upper} + IF_{lower})^{123}$$

ID = item discrimination for an individual item

IF_{upper} = item facility for the upper group on the whole test

IF_{lower} = item facility for the lower group on the whole test

Item discrimination of the test is calculated using SPSS. If item discrimination (ID) value is less than 0, item is very bad (must be revised). Then, if item discrimination (ID) value is between 0 until 0.19, item is bad. Then, if item

[123] Ibid, 69.

discrimination (ID) value is between 0.20-0.29, item is medium. Then, if item discrimination (ID) value is between 0.30-0.39, item is good. If item discrimination (ID) is more than 0.40, item is very good.[124] The results are:

Table 1.12 Item Discrimination Analysis

No. Item	Item Discrimination	Interpretation
1	0,137	Poor
2	0,296	Medium
3	0,184	Medium
4	0,444	Very good
5	0,395	Good
6	0,382	Good
7	0,331	Good
8	0,501	Very good
9	0,294	Medium
10	0,482	Very good
11	0,328	Medium
12	-0,020	Very bad
13	0,459	Very good
14	0,473	Very good
15	0,336	Medium
16	0,096	Poor
17	0,356	Medium

[124] Wahyuni and Ibrahim, *Asesmen Pembelajaran Bahasa*, 136.

18	0,356	Medium
19	0,524	Very good
20	0,126	Poor
21	-0,081	Very bad
22	0,589	Very good
23	0,486	Very good
24	0,165	Poor
25	0,335	Good
26	0,437	Very good
27	0,271	Medium
28	0,232	Medium
29	0,219	Medium
30	0,336	Good
31	0,304	Good
32	0,428	Very good
33	0,501	Very good
34	0,606	Very good
35	0,388	Good
36	0,529	Very good
37	0,524	Very good
38	0,393	Good
39	0,224	Medium
40	0,081	Poor
41	0,410	Very good
42	0,443	Very good
43	0,684	Very good
44	0,569	Very good
45	0,315	Good
46	0,433	Very good
47	0,591	Very good
48	0,481	Very good

49	0,316	Good
50	0,308	Good

From the table above, there are 2 very poor items, they are number: 12 and 21. Then, there are 5 poor items, they are number: 1,16,20,24, and 40. Then, there are 11 medium items, they are: 2,3,9,11,15,17,18,27,28,29, and 39. Then, there are 11 good items, they are number: 5,6,7,25,30,31,35,38,45,49, and 50. Then, there are 21 very good items, they are number: 4,8,10,13,14,19,22,23,26,32,33, 34,36,37,41,42,43, 44,46,47, and 48.

5. Item Distractor

The formula which is used to calculate item distractor is:

$$IDs = 5\% \times n_{ts}$$

IDs = percentage

n_{ts} = total of the students who took the test.

Item distractor of the test can be said as good, if 5% of the participants choose the distractors. The participants of the test are 298 people. 5% x 298 = 14,9. The results are:

Table 1.13 Distractor Items Analysis Results

NO	ANSWER OPTIONS				ANSWER KEYS	INTERPRETATION
	A	**B**	**C**	**D**		
1	275	7	7	7	**A**	Item B,C, and D are not good and they need to be revised
2	17	23	239	17	**C**	All items are good
3	31	237	9	19	**B**	All items are good
4	67	81	63	85	**D**	All items are good
5	58	91	65	82	**D**	All items are good
6	245	22	18	11	**A**	Item D is not good and it need to be revised
7	44	131	74	47	**A**	All items are good
8	10	46	42	198	**D**	Item A is not good, and it need to be

						revised
9	200	47	30	19	**A**	All items are good
10	211	35	26	24	**A**	All items are good
11	15	248	18	15	**B**	All items are good
12	79	86	63	68	**C**	All items are good
13	30	44	16	206	**D**	All items are good
14	49	146	53	48	**B**	All items are good
15	19	250	11	16	**B**	Item C is not good, and it need to be revised
16	78	86	89	43	**A**	All items are good
17	42	114	81	59	**A**	All items are good
18	38	13	231	14	**C**	Item B is not good, and it need to be revised
19	219	26	28	23	**A**	All items are good
20	62	108	71	55	**D**	All items are good
21	73	55	105	63	**B**	All items are good
22	36	209	39	12	**B**	Item D is not good, and it need to be revised
23	222	27	23	24	**A**	All items are good
24	53	48	126	69	**B**	All items are good
25	82	58	88	68	**B**	All items are good
26	243	14	27	12	**A**	Item B and D are not good, and it need to be revised
27	96	35	109	56	**B**	All items are good
28	60	77	86	73	**D**	All items are good

29	65	91	90	50	**D**	All items are good
30	235	19	24	18	**A**	All items are good
31	72	43	103	78	**B**	All items are good
32	60	81	96	59	**A**	All items are good
33	64	85	84	63	**A**	All items are good
34	205	46	36	9	**A**	Item D is not good, and it need to be revised
35	26	64	62	144	**D**	All items are good
36	22	199	39	36	**B**	All items are good
37	27	90	90	89	**A**	All items are good
38	15	37	226	18	**C**	All items are good
39	85	69	91	51	**A**	All items are good
40	87	91	72	46	**C**	All items are good
41	84	37	105	70	**B**	All items are good
42	30	27	206	33	**C**	All items are good
43	54	50	146	46	**C**	All items are good
44	16	36	31	213	**D**	All items are good
45	40	85	104	67	**A**	All items are good
46	10	29	30	227	**D**	Item A is not good, and it need to be revised
47	99	64	71	62	**A**	All items are good
48	24	21	235	16	**C**	All items are good
49	67	82	68	79	**D**	All items are good
50	74	90	81	51	**D**	All items are good

From the table above, the results there are 9 questions that some items need to be revised. They

are number: 1 (item B,C,D need to be revised), 6 (item D needs to be revised), 8 (item A needs to be revised), 15 (item C needs to be revised), 18 (item B needs to be revised), 22 (item D needs to be revised), 26 (item B and D need to be revised), 34 (item D needs to be revised), and 46 (item A needs to be revised).

D. Interpretation and Discussion

The purpose of this study is giving the overview of English Try-out test in MTs N 2 Ponorogo by conducting item analysis of try-out test (Computer-Based Test) administered to the student of ninth grade. Items analysis is conducted by calculating item validity, item reliability, item difficulty, item discrimination, and item distractor.

Referring from the explanation above, the validity of the test is analyzed using IBM SPSS program version 23. The result shows that there are 2 numbers with very low validity, 23 with low validity, 19 numbers with medium validity, and 2 numbers with high validity. Therefore, the researcher concluded that the test has low validity.

Based on the explanation above, the reliability of the test is analyzed using SPSS program version 23. The result shows alpha scale 0,867 and r_{table} is 0,113. Therefore, based on the degree of reliability, the researcher concludes that the test has very high reliability.

From the result of item difficulty analysis, there are 16 easy test items, they are number: 1,2,3,6,10,11, 15,18,19,23,26,30,38,44,46, and 48. Then, there are 11

medium test items, they are number: 8,9,13,14,22,34, 35,36, 42,43, and 47. Then, there are 23 difficult test items, they are number: 4,5,7,12,16,17,20,21,24,25,27, 28,29,31,32,33,37,39,40,41,45,49, and 50. Therefore, the researcher concludes that the level of difficulty of the test is difficult.

From the result of item discrimination analysis, there are 2 very bad items, they are number: 12 and 21. Then, there are 5 poor items, they are number: 1,16,20,24, and 40. Then, there are 11 medium items, they are: 2,3,9,11,15,17,18,27,28,29, and 39. Then there are 11 good items, they are number: 5,6,7,25,30,31,35,38,45,49, and 50. Then, there are 21 very good items, they are number: 4,8,10,13,14,19,22,23,26,32,33,34,36,37,41,42,43,44,46

,47, and 48. As a result, the researcher concludes that the test has very good item discrimination.

From the result of item distractor analysis, there are 9 questions that some items need to be revised. They are number:1 (item B,C,D need to be revised), 6 (item D needs to be revised), 8 (item A needs to be revised), 15 (item C needs to be revised), 18 (item B needs to be revised), 22 (item D needs to be revised), 26 (item B and D need to be revised), 34 (item D need to be revised), and 46 (item A need to be revised). Therefore, the researcher concludes that the test has good criteria on item distractor.

Finally, the teachers must analyze the result of the test. Because, teachers will know about the result of the test in terms of validity, reliability, item difficulty, item

discrimination, and item distractor and they can know what part that needs to be revised.

CHAPTER V

CLOSING

This chapter discusses about conclusion and recommendation. The conclusion comes from the result of the research. And the recommendation presents the suggestion for some aspects.

A. Conclusion

Based on the test items analysis consisting of item validity, item reliability, item difficulty, item discrimination, and item distractor on English Try-out Test (Computer-Based Test) at MTs N 2 Ponorogo using IBM SPSS program version 23 can be concluded as follows:

1. The finding shows that try-out test has low validity, with r_{table} 0,113 and significance level 5%. It is

drawn from the result of analysis 4% of the item test has very low validity, 46% has low validity, 38% has medium validity, and 4 % has high validity.

2. The result of item analysis reveals that the test has very high reliability, with the alpha scale 0,867 and r_{table} is 0,113.

3. The finding shows that the level of difficulty of the test is difficult. There are 32% easy test items, 22% medium items, and 46% difficult items. In which, there are 16 easy items, 11 medium test items, and 23 difficult test items.

4. The result of item analysis shows that this test has very good item discrimination where 21 of 50 items has very good item discrimination, there are 4%

very poor items, 10% poor items, 22% medium items, 22% good items, and 42% very good items.

5. The finding shows that the test has good criteria on item distractor. There are 9 questions that need to be revised. They are number: 1 (item B,C,D need to be revised), 6 (item D needs to be revised), 8 (item A needs to be revised), 15 (item C needs to be revised), 18 (item B needs to be revised), 22 (item D needs to be revised), 26 (item B and D need to be revised), 34 (item D needs to be revised), and 46 (item A needs to be revised).

B. Recommendation

Based on the result of the research, the researcher offers some recommendations to the test maker. First, test makers need to analyze the result of the test to know the quality of the test. Second, test makers need to

revise test items which need to be revised. And the last, test makers need to give students motivation to do the test seriously although the test is try-out test.

BIBLIOGRAPHY

Alabi, A.T., A. O. Issa, and R. A. Oyekunle, "The Use of Computer Based Testing Method for the Conduct of Examinations at the University of Ilorin," *International Journal of Learning & Development*, 3, (April-May, 2012).

Alderson, J. Charles and Caroline Clapam. *Language Test Construction and Evaluation*. Trumpington: Cambridge University Press, 1995.

Arikunto, Suharsimi. *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara, 2012.

Ary, Donald, Lucy Cheser Jacobs, and Christine K. Sorensen. *Introduction to Research in Education.* Belmont: Wadsworth, 2010.

Balnaves, Mark and Peter Caputi. *Introduction to Quantitative Research Methods: An investigative Approach*. London: Sage Publicaions, 2001.

Brown, H. Douglas. *Language Assessment Principles and Classroom Practices.* New York: Longman, 2000.

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

Brown, James Dean. *Testing in Language Programs: A Comprehensive Guide to English Language*

Assessment. New York: The Mac-Graw Hills Companies, 2005.

Cohen, Louis, Lawrence Manion, and Keith Morrison. *Research Methods in Education.* London: Taylor & Francis e-Library, 2005.

Creswell, Jhon W.. *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research.* Boston: Pearson Education Inc., 2012.

Dirman, CD. and Cicih Juarsih. *Penilaian dan Evaluasi*. Jakarta: Rineka Cipta, 2014.

Elfiza, Yulia, Rusman, and M. Nasir," Hubungan antara Hasil Uji Kognitif Try Out Ujian Nasional (UN) dengan Hasil Ujian Nasional (UN) Mata Pelajaran Kimia SMA Kota Banda Aceh Tahun Ajaran 2014/2015", *Jurnal Ilmiah Mahasiswa Pendidikan Kimia*, 1, (2016).

Gronlund, Norman E.. *Constructing Achievement Test*. New Jersey: Prentice Hall Inc., 1977.

Hening, Grant. *A Guide to Language Testing: Developtment, Evaluation, and Research*. Foreign Language and Research Press, 2001.

Hughes, Arthur. *Testing for Language Teachers*. Cambridge: Cambridge University Press, 2003.

Kothari, C.R.. *Research Methodology Methods and Techniques.* New Delhi: New Age Publishers, 2004.

Neuman, W. Lawrence. *Basic of Social Research: Qualitative and Quantitative Approach*. Boston: Pearson Education, 2007.

Noveria, Anita, ”Item Analysis on the Validity and the Reliability of the English Summative Test for the first-year students at MA Madani Alauddin Pao-pao”, *ISERD International Conference*, (March, 2018).

Noorrachma Chandra Novianti, “*Test Item Analysis of the Final Examination on Economics Subject in Grade XII IPS SMA Negeri 1 Wonosari Academic Year 2014/2015*”,Thesis (Yogyakarta: Yogyakarta State University, 2015).

Purwanto, M. Ngalim. *Prinsip-prinsip dan Teknik Evaluasi Pembelajaran*. Jakarta: Remaja Rosdakarya, 2013.

Ratnaningsih, Arrynda, “*Analysis Kualitas Soal-Soal Try Out Ujian Nasional Mata Pelajaran IPA SMP di Kabupaten Banjarnegara*”, Thesis (Semarang: Semarang Stated University, 2012).

Russell, Michael K. and Peter W. Airasian. *Classroom Assessment: Concept and Application*. New York: McGraw Hill, 2012.

Sakti, Rafikasari Risqi, *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 (Ponorogo: Stated institute of Islamic Studies Ponorogo, 2019).

Sukardi, H.M.. *Evaluasi Pendidikan: Prinsip dan Operasionalnya*. Jakarta: BumiAksara, 2011.

Toksoz, Sibel and Ayse Ertunc, "Item Analysis of Multiple-Choice Exam", *Advances in Language and Literary Studies*, 8, (December 2017).

Wahyuni, Sri and Abd. Syukur Ibrahim. *Asesmen Pembelajaran Bahasa*. Bandung: PT Refika Aditama, 2012.

Widyaningrum, Retno. *Statistika.* Depok: Pustaka Felicha, 2015.

Yasar, Sefik and Asli Gundogan Cogenli, "Determining Validity and Reliability of Data Gathering Instruments used by Program Evaluation Studies in Turkey", *Procedia Social and Behavioral Sciences*, (2014).