Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 14, Issue 1, January 2023:68-89

The Characteristics of the First Semester Final Test Indonesian Class 7

Suwarto Suwarto[a]

[a] Prof. Dr., Veteran Bangun Nusantara University, Sukoharjo, Indonesia,
e-mail: suwartowarto@yahoo.com, ORCID ID: 0000-0002-7925-8017.

Abstract

The purpose of this research is to describe: the characteristics of the Indonesian language test based on classical test theory and modern test theory. The research design is quantitative and descriptive. The object of the research is the Indonesian language achievement test, Indonesian language teachers, school principals, and vice principals. The data was obtained from 280 student responses to all answer sheets of 7th grade students of SMP IT MTA Karanganyar as the population of this study. The answer keys to the Indonesian language questions and one package of Indonesian questions (40 multiple choice items) were obtained from the Indonesian language teacher. Research techniques with interviews and documentaries. Data analysis was carried out using the Quest program. The results of the study: (1) Characteristics of the Indonesian language test based on classical test theory: (a) The category of item difficulty in the percentage of easy: medium: difficult = 55%: 37.5%: 7.5%; (b) The item discriminatory category in poor percentage: adequate: good: very good = 10%: 30%: 25%: 35%; (c) Distractor function in ineffective percentage: effective = 46.70%: 53.30%; (d) Test reliability 0.990; and (e) The validity of the test content is met. (2) Characteristics of Indonesian test based on modern test theory: (a) Threshold category of Indonesian test in percentage of very difficult: difficult: medium: easy: very easy = 7.5%:22,5%:35%:25%: 10%; (b) The percentage of match between Indonesian test items and the Racsh model is 100%.

Keywords: The characteristic tests, item difficulty, item discrimination, reliability.

Endonezya Dili 7. Sınıf Birinci Dönem Son Testinin Özellikleri

Öz

Araştırmanın amacı, klasik test teorisine ve modern test teorisine dayalı Endonezya dili testinin özelliklerini açıklamaktır. Araştırma tasarımı nicel ve tanımlayıcıdır. Araştırmanın amacı Endonezya dili başarı testi, Endonezya dili öğretmenleri, okul müdürleri ve okul müdür yardımcılarıdır. Veriler, bu çalışmanın evrenini oluşturan SMP IT MTA Karanganyar'daki 7. sınıf öğrencilerinin tüm cevap kağıtlarına 280 öğrencinin verdiği yanıtlardan elde edilmiştir. Endonezce soruların cevap anahtarları ve Endonezyaca sorulardan oluşan bir paket (40 çoktan seçmeli madde) Endonezya dili öğretmeninden temin edildi. Röportaj ve belgesellerle araştırma teknikleri. Veri analizi Quest programı kullanılarak yapıldı. Çalışmanın sonuçları: (1) Klasik test teorisine dayalı Endonezya testinin özellikleri: (a) Kolay: orta: zor = %55: %37,5: %7,5 oranında kategori madde zorluk seviyesi; (b) Tane diferansiyel gücünün yüzde olarak kategorisi kötü: yeterli: iyi: çok iyi = %10: %30: %25: %35; (c) Etkisiz yüzdede çeldirici

işlevi: etkin = %46,70: %53,30; (d) Test güvenilirliği 0.990; ve (e) Test içeriğinin geçerliliğinin karşılanması. (2) Modern test teorisine dayalı Endonezya testinin özellikleri: (a) Endonezya testinin çok zor: zor: orta: kolay: çok kolay = %7,5:22,5:35:25% açısından eşik kategorisi: %10; (b) Endonezya dili testi öğeleri ile Racsh modeli arasındaki eşleşme yüzdesi %100'dür.

Anahtar Sözcükler: Karakteristik testler, madde güçlüğü, madde ayırt ediciliği, güvenirlik**.**

1. Introduction

The quality of education is a problem that the government continues to strive to improve. Improving education is one of the national education development priorities. Educational development is also an effort to increase human resources. To achieve quality human resources, education must be of high quality. However, to achieve quality education, there are still many obstacles faced by the world of education in Indonesia, including the quality of educators and education staff, students, infrastructure, learning processes including evaluation and assessment processes.

Based on government regulation number 19 of 2005 concerning National Education Standards Article 66 paragraph 1 it is stated that the assessment of learning outcomes by the government aims to assess the competency achievement of graduates nationally in certain subjects in the science and technology subject group and is carried out in the form of a national exam. Assessment is an important component in the education system to determine the progress and level of achievement of learning outcomes. Assessment requires good data. One source of data is the measurement results. This measurement activity is usually done through a test. The quality of a good test will determine the quality of the formulation of the assessment results, and will further determine the quality of various educational decisions or policies. The learning achievement test aims to measure a person's ability or achievement after undergoing the learning process. Tests like this are very important to be carried out by teachers, schools and educational institutions to find out how far students have achieved the expected learning goals. Test results can be used by teachers, schools or other educational institutions to make decisions or provide feedback for improving the teaching and learning process. The test is an evaluation tool that has an important role in measuring the teaching and learning process. Shomami (2014) supports that tests are samples of knowledge and need to be good representations of them. That is, what is tested is only a sample of behavior or knowledge, not the whole or behavior taught by the teacher and learned by students because it is impossible to measure all students' abilities. A test should be properly regulated so that it can be used effectively. To be said to be a good test must meet the characteristics of a good test, namely; validity, reliability, practicality, level of difficulty, differentiability and especially for multiple choice tests have effective distractors for each item. Tests are used by many schools across the country to gather information about students' abilities and knowledge.

To determine the quality of the test, the test items must be analyzed. Test questions must also be prepared properly, because the test results will be affected by the quality of the test. The quality of the test is influenced by the quality of each item. Teachers should focus on the quality of the test items, so it is very important for teachers to do item analysis. Because by analyzing the items, the teacher can identify the quality of each item, know which items match the criteria, which items should be eliminated, and which items should be revised. To assess student achievement, the teacher usually gives several questions to students in the form of tests. The teacher can carry it out after each material chapter is finished or at the end of the semester, the test is called an achievement test. Achievement

tests are a systematic procedure for determining the number of students who have learned. There are two kinds of achievement tests; formative tests and summative tests.

A summative test is an assessment activity that produces scores or numbers which are then used as a decision on student achievement. This test is carried out if the unit of learning experience or all of the subject matter has been completed. A summative test is used to determine the award classification at the end of a course or program. Then, formative tests are for monitoring student learning to provide ongoing feedback that can be used by instructors to improve their teaching and by students to improve their learning. More specifically, formative tests: help students identify their strengths and weaknesses and target areas that need work, and help faculty recognize where students are struggling and address issues promptly. In this study, researchers used summative test items as data sources. The researcher analyzed the test items to assess test quality such as validity, reliability and others. The results of this analysis are used for the following school year as a consideration for a teacher in explaining material, if possible, not being mastered by students.

Based on the background presented and the reasons above, the researcher formulates a research question: what are the characteristics of the final test of the first semester of Indonesian based on classical test theory and based on modern test theory? Questions related to the characteristics of the final test of the first semester of Indonesian based on classical test theory are as follows: (1). What is the difficulty level of the Indonesian language test items; (2). How is the different power of the Indonesian language test items; (3). How does the Indonesian language test distractor function; (4). How is the reliability of the Indonesian language test; (5). What is the content validity of the Indonesian language test; Questions based on modern test theory as follows: (6). How is the Indonesian test Threshold category; and (7). How do the Indonesian language test items match the Racsh Model.

There have been a number of studies conducted in the past related to this research. These studies analyzed the characteristics of the test, namely, item difficulty level, item discriminatory power, alternative answers, and reliability. First, the researchers analyzed item difficulty (p). Singh, Kariwal, Gupta, & Shrotriya (2014) analyzed the quality of multiple-choice questions through item analysis. He found 11 (5%) items belonging to the moderate category with a range of 30% - 70%, 9 (45%) items belonging to the easy category with a range of p> 70%, and no items belonging to the difficult category with a range of p <30 %. Chauhan, Ratrhod, Chauhan, & Rameshbhai (2013) calculated the difficulty level of each item in a multiple-choice test on the subject of anatomy. They found 35 out of 65 items to be in the acceptable range (30-50% or 60-70%), 3 out of 65 items to be in the difficult category (p <30%), 12 out of 65 items to be in the easy category (p> 70%), and 15 out of 65 items included in the ideal quality (50-60%). Most of the items are acceptable item difficulty levels. Suruchi & Rana (2012) analyzed the difficulty of the biology achievement test items. They found 1 item out of 120 items included in the difficult category with a range below 0.20, 18 items out of 120 items included in the good category with a range of 0.20-0.50, 94 items out of 120 items included in the best category with a range 0.50-0.80, and 7 out of 120 items belonging to the very easy category with a range above 0.80. Thus, they determined that one difficult item and seven easy items should be rejected for the final achievement test draft. Kolte (2015) found 4 difficult items with a p range <30%, 26 acceptable items with a range of 30-70%, and 10 easy items with a p>70% range. Sa'adah (2017) analyzed the quality of the test items for the English mid-semester exam. He found 18 items (72%) as ideal categories with a range of about 0.62, 2 items (8%) as easy categories with a range of p> 0.90 and 5 items (20%) as

difficult categories with a range of p <0. 20. Saputra (2015) compared test quality for the second semester English test between SMP N 1 Semarang and SMP Kesatrian 2 Semarang. He found 31 easy items, 15 medium items, and 4 difficult items from SMP N 1 Semarang, while in SMP Kesatrian 2 Semarang there were 36 easy items and 14 moderate items.

The researchers above analyzed the level of item difficulty through certain formulas such as item difficulty level (p) or prop-correct, including that they analyzed it manually, however, there were researchers who used programs to analyze it, for example, Mulianah & Hidayat (2013) used programs Iteman version 3.00 to analyze computer-based test items including item difficulty level. In addition, each researcher chooses a particular category of theory to determine the quality of the items. Second, the researchers analyzed the differential power of the items (D). This is one of the items analyzes that can be calculated manually or through computer software such as SPSS, Microsoft Excel, Anates and Iteman programs. For example, Boopathiraj & Chellamani (2013) have analyzed the differential power of items using the separation method between the upper and lower groups whose scores are entered in Microsoft Excel. Ainol Madziah Zubairi (2006) used SPSS and Bigsteps to analyze the characteristics of the items which are the difficulty level of the items and the differential power of the items. Another example, Raharja (2014) analyzed grain discrepancy with Anates V4. In his study, there was no very good grain discriminant category. There are only 8 items in the good category, 13 items in the sufficient category, and 28 items in the bad category. Therefore, bad items should be removed, and sufficient items should be revised. In addition, a bad item indicates that the item cannot differentiate between students who have mastered the competency and students who have not mastered the competency.

Third, the researchers analyzed the distractors. Putri (2015) analyzed it with the Iteman program version 3.00, however, for reliability, item difficulty level, and item discriminatory power, she analyzed it manually. It shows that he doesn't know that Iteman program version 3.00 can analyze everything using Iteman program version 3.00. This has been done by Rusmiana (2015). He used the Iteman program version 3.00 to determine the characteristics of a good test. The final characteristic of a test is reliability. A test is said to be reliable if the test is consistent from time to time to produce the same score. Many studies were conducted in the past that found reliability using the Kuder-Richardson 20/21 formula (KR-20/KR-21). The researchers who found the reliability of the test were Sugianto (2017), Haryudin (2015), Bernasela (2014), and Pascual (2016). Pascual (2016) described that the English achievement test for ESL students in the Northern Philippines is reliable. However, there is a researcher, Hidayati (2009) who found moderate reliability in the mid-semester English test for seventh graders of SMPN 33 Semarang. The difference between the above studies and this research is the data analysis technique used by the Iteman version 3.00 program to analyze and reveal the difficulty level of the items, the differential power of the items, the distractors and even the answer keys. However, it can also be a similarity because Rusmiana (2015) also uses the Iteman program version 3.00 to analyze test characteristics. The difference between Rusmiana's research and this research is the object of research. Rusmiana's study analyzes accounting theory for vocational education, while the object of this research is a summative test which is the final test of the first semester of Indonesian for seventh grade students. Another similarity between the above studies and this research is the descriptive quantitative approach. Most of these studies use descriptive quantitative methods.

Classical Test Theory and Modern Test Theory

Measurement in education includes measuring the ability of test takers and measuring the characteristics of the measuring instruments used. There are two measurement theories that are still being developed, namely the classical test theory and the modern test theory. Classical test theory is also called Classical Test Theory (CTT), while modern test theory is also called Item Response Theory (IRT). Classical test theory is based on assumptions. The following assumptions are summarized from classic test theory by Allen & Yen (1979). The theory's first assumption is that a test taker's observed score is the sum of his actual score and his error score. X = T+E; where: X is the achievement score; the T score is correct and the error score is E. The error score is a discrepancy in the score obtained from the actual situation and occurs randomly. The second assumption in classical test theory is that the population mean of the observed scores is the same independent value as the true score for each test taker on the same test. [E(X)=T] This means that the actual score is the average value of the theoretical acquisition score if repeated measurements are made using the same measuring instrument. The third assumption; the true score and the achieved error score in a population on a test are not correlated [ρ_{ET}=0]. This insurance provides an understanding that there is no correlation between the actual score and the error score. A test taker who has a high correct score does not necessarily have a high error value, so does a participant who has a low correct score does not necessarily have a high error value. Fourth Assumption; the correlation between the error scores on the first test and the error scores on the second test is zero; [ρ_{E1E2}=0]. This means that test takers who have a high error score in the first test may not necessarily get an error score in the second test. This provides an understanding that the range of implementation of the first and second tests can be influenced by the situation and objective conditions of the test takers. Fifth assumption; on tests that measure the same attribute, the error score on the first test does not correlate with the true score on the second test. These assumptions (first to fifth) provide a very simple interpretation of test scores. So that the characteristics of the test and even the test takers are based on the test results in the group. And systematic error cannot be called measurement error (Allen & Yen, 1979). Sixth assumption; two sets of tests that measure the same trait, resulting in an acquisition score X, and the true scores are T_1 and T_2, and the variance of the scores σ^2_1 and σ^2_2, these two sets of tests are said to be parallel if $T_1 = T_2$ and $\sigma^2_1 = \sigma^2_2$ and satisfy the first to fifth assumptions. Seventh assumption; If two sets of tests are intended to measure the same trait and meet the first to fifth assumptions, the tests are said to be equivalent if in each population the test takers score from the first test (X_1) is the same as the score on the second test added a constant (C), [$X_1 = X_2 + C_{12}$].

To overcome the weaknesses that exist in the classical theory, measurement experts try to find alternatives. The desired model must have the following characteristics: (1) the characteristics of the items do not depend on the group of test takers subjected to the questions, (2) the value indicating the ability of the test takers does not depend on the test, (3) the model is stated in levels. items, not in the level of the test, (4) the level model does not require parallel tests to calculate the reliability coefficient, and (5) the model provides an appropriate measure of each ability score (Hambleton, Swaminathan, & Rogers, 1991). An alternative model that can have these characteristics is a measurement model called modern test theory or Item Response Theory.

Modern test theory was developed by measurement experts in the fields of psychology and education as an effort to minimize the deficiencies that exist in classical test theory. As in classical theory, modern

test theory is also based on basic postulates. There are two basic postulates of modern test theory (Hambleton et al., 1991), namely: (1) the test taker's work on an item can be predicted from a type of factor called trait, latent trait, or ability; (2) the relationship between the test taker's work on the test item and the underlying traits can be explained by a monotonic increasing function called the item characteristic function (item characteristic function or item characteristic curve-ICC). This function explains, if the ability level increases, then the probability of answering correctly on a test item also increases.

As in classical test theory, modern test theory also contains underlying assumptions, namely: (a) Unidimensionality, (b) Local independence, and (c) Item characteristic functions express the correct relationship between unobserved variables (namely ability). with the observed variables. (i.e. item response) (Hambleton et al., 1991); (Sumadi, 2005). The assumptions of unidimensionality and local independence can be explained as follows.

The unidimensionality assumption states that only one ability is measured by a set of items in a test. This assumption is difficult to fulfill in practice, because many factors can affect the results of a test. These factors include the level of motivation, anxiety, ability to work quickly, and other cognitive skills beyond the ability as measured by a set of items in a test. What is meant by unidimensionality in this case is the existence of a dominant factor that influences the results of a test. This dominant factor is called ability as measured by the test. The assumption of local independence states that the ability attitude that influences a test is constant, so the test taker's response to each pair of items is statistically independent. In other words, the assumption of local independence states that there is no correlation between test takers' responses to different items. This also means that the ability stated in the model is the only factor that influences the test taker's response to the items. Modern test theory uses 1 logistic parameter (1PL), namely item difficulty level is defined as a score on a test taker's ability scale which has a probability of 0.50 to answer correctly on a particular item (Hambleton & Rogers, 1988). The one-parameter logistic model is often referred to as the Rasch model, in honor of its discoverer (Hambleton et al., 1991). (Lord, 1980) developed a two-parameter logistic (2PL) model based on the Ogive normal distribution. Lord is seen as the first to develop a two-parameter logistic item response model (Hambleton et al., 1991). Then it developed again into three logistic parameters (3PL).

The classical test theory contains various advantages and disadvantages. The advantages of classical test theory include: (1) using simple concepts to determine the ability of test takers, (2) using simple concepts in calculating the validity and reliability coefficients of tests and calculating item parameter values, (3) can be used in small samples., for example at the grade level, (4) has long been used in practical measurement and testing, so it has been known and understood by most people involved in or related to the world of education and psychology. Meanwhile, as previously mentioned, the weaknesses of the classical test theory include: (1) the ability of the test takers expressed in discrete variables, and (2) the magnitude of the validity and reliability coefficients of a test and the parameter values of an item. depending on the test taker subjected to the test.

Because the emergence of modern test theory is intended to cover the weaknesses in classical test theory, the advantages of modern test theory include: (1) the theoretical basis is better than classical test theory, (2) the ability of test takers is expressed in continuous variables, (3) there is no need for parallel tests to calculate the reliability coefficient (which in modern test theory is called the

information function), and (4) the magnitude of the reliability coefficient of a test and the parameter values of an item do not depend on the test takers subjected to the test. . However, the use of modern test theory contains several weaknesses, including: (1) it requires a large sample to be able to produce stable parameters, so that modern theoretical concepts cannot be applied at the class level, (2) software requires a reliable computer program to be able to perform accurate parameter estimation, and (3) its existence has not been accepted by the majority of people working in the world of education and psychology, especially in Indonesia.

2.Methodology

Research Design

In this study, the researcher revealed the characteristics of the first semester final test of the Indonesian language by analyzing the test items and the reliability of the test, so the researcher used a descriptive quantitative approach. This study uses descriptive analysis because it is intended to reveal the characteristics of the test at the end of the first semester's test of Indonesian for seventh grade students. Researchers used quantitative research because the numerical data were analyzed statistically with the Quest program.

Research Place

This research was conducted at MTA Karanganyar IT Middle School located on Jalan Karanganyar Tawangmangu at the end of April 2022 after the first semester final exams for Indonesian were held for seventh graders.

Research subject

The subjects of this research are research participants who participate in certain research by becoming research targets. However, researchers used student answer sheets to become the target of this study. There are two kinds as follows: The population of this research is all seventh-grade students' answer sheets. The population of this study was 280 student answer sheets which can be seen in Table 1.

Table 1 *Research Population*

Class	Male	Fimale	Total
7A	15	20	35
7B	15	20	35
7C	19	16	35
7D	19	16	35
7E	15	20	35
7F	19	16	35
7G	15	20	35
7H	19	16	35
Total			280

According to Suwarto (2018), the sample represents a population. In selecting the sample, the researcher used random sampling. The sample takes 10-15% or 20-25% of the population. The greater

this percentage the better, so the researcher took 100% of the 280 student answer sheets. So the entire population is taken data.

Research Object

The object of this study was a multiple-choice test of the 7th grade Indonesian first semester final exam at SMP IT MTA Karanganyar in the 2021/2022 academic year.

Research variable

The variables of this study were item analysis (item difficulty level and item differentiating power), alternative analysis (answer keys and distractors), and the reliability of the first semester final test of Indonesian for seventh grade students at SMP IT MTA Karanganyar in the 2021/2022 academic year.

Data collection technique

This study used two data collection techniques, namely interviews and documents. Interviews were conducted by researchers to collect data. First, the researcher asked permission to conduct research in schools from the principal and administration. Second, the researcher asked one of the Indonesian teachers to get information about the school's program curriculum and data on all seventh-grade students, then, confirmed the time for the researcher to collect the data (first Indonesian final test paper, student answer sheets). Then, the chairman of the exam committee was also interviewed to obtain information about the procedure for making the final Indonesian first semester test made by Indonesian teachers who are members of the Indonesian Language MGMP in Karanganyar Regency, and the researcher asked whether the test had been analyzed and tried before. In addition, the researcher asked him the answer key. Documents consist of: Indonesian language first semester final exam sheets, answer keys, and student answer sheets. From this answer sheet, they will be analyzed from each question about the level of difficulty of the items, the differential power of the items, distractors, and reliability.

Data analysis technique

The multiple-choice tests and answers to the first semester final exam in Indonesian at SMP IT MTA Karanganyar in the 2021/2022 academic year were analyzed to find out whether each item was easy, moderate, or difficult for students. For grain discriminatory power, is the quality of each grain poor, fair, good, very good? Does each item have a functioning distractor or not. Not only that but also to find out whether the multiple-choice test is reliable or not. For this purpose, all will be analyzed using the Quest program.

Item Analysis Based on Classical Test Theory

Item Difficulty

To find the item difficulty of each test item, the following formula:

$$p = \sum B \diagup N \quad \text{.....(1)}$$

Where:

P = proportion of correct

$\sum$ B= the number of correct answers

N= the number of respondents. (Lababa, 2018).

Item difficulty levels can be classified into three, namely: easy, medium, and difficult. According to (S Suwarto, 2017); (Suwarto Suwarto, 2017); (Mutaqi, 2007), the item difficulty level categories are as follows:

Table 2.

The Category of the Item Difficulty

P = The item difficulty	Category
P > 0.700	Easy
0.300 ≤ p ≤ 0.700	Moderate
P < 0.300	Difficult

P = Prop. Correct is the proportion of students answering correctly. Item difficulty level close to 0 or 1 indicates the item is too difficult or too easy for students (Suwarto Suwarto, 2021).

Item Discrimination

The different power of the test items can be found using the Point-Biserial Correlation Formula. To find out the differential power of each item, use the point-biserial correlation formula with the following formula:

$$r_{pbi} = \frac{M_p - M_t}{S_T}\sqrt{\frac{p}{q}} \quad \text{.....(2)}$$

Where:

r_{pbi} = point biserial correlation coefficient

M_p = the mean criterion score for those who answer the item correctly

M_t = the mean criterion of total score

S_T = standard deviation of total score

p = *proportion of correct*

q = *proportion of false* (q = 1 – p) (Crocker, L., and Algina, 1986).

(Suwarto Suwarto, 2021); (Suwarto, 2018) state that point-biserial correlation is a bivariate correlation technique. To use this technique, variable 1 is discrete data (dichotomous data), and variable 2 is continuum data (interval data). This technique is usually used to calculate the differential power of items by correlating between the item scores and the total score. The point-biserial correlation

coefficient (r_{pbi}) is a statistical measure used to estimate the degree of relationship between the nominal dichotomous scale and the interval scale (J. D. Brown, 2001).

The grain discriminatory power can be classified into four, namely: bad, fair, good, and very good. Bad category items should be removed. Items in the sufficient category must be revised, however, for items that are in the good and very good categories are accepted and stored in the question bank (Mulianah & Hidayat, 2013); (Suwarto Suwarto, 2021); (S Suwarto, 2017).

Table 3. *The Category of the Item Discrimination*

r_{pbis} = Item Discrimination	Category
$r_{pbis} \leq 0.200$	Bad
$0.200 < r_{pbis} \leq 0.400$	Sufficient
$0.400 < r_{pbis} \leq 0.700$	Good
$r_{pbiss} > 0.700$	Very Good

Alternatives Analysis

The distractor is said to be effective if it is selected at least 5% (0.050) of the respondents. The distractor is said to be ineffective if less than 5% of respondents are selected (Suwarto Suwarto, 2016). All distractors that are not effective should be revised. (Lababa, 2018) states that distractors that do not meet the criteria must be replaced or revised with other distractors that may be more interesting and confusing for students.

Reliability Tests

The researcher determines reliability with the Alpha Crobanch (α) formula because the Quest program, which can be seen on the last page of the Quest program output, uses Alpha to show reliability. Not only does the Quest program use this formula, but also the final exam for the first semester of Indonesian is an instrument where the answer is a scale (dichotomous). The answer only has two answers which are the correct answer (score 1) and the wrong answer (score 0). This formula can be used to calculate the dichotomous scale (Suwarto Suwarto, 2021). The Cronbach Alpha reliability coefficient formula (α):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s^2{}_i}{s^2{}_x}\right) \quad \text{.....(3)}$$

Where:

$\alpha = Alpha - Cronbach$

k = the number of subtests.

$\sum s^2_i$ = sum of all test variant items

s^2_x = variance of the total test score

(Suwarto Suwarto, 2021)

The reliability index ranges from 0-1. A test is said reliable if the reliability index upper 0.700. The highest reliability coefficient of a test is close to index 1. It indicates that a test has perfect reliability (Suwarto Suwarto, 2021)(Suwarto Suwarto, 2016).

Content Validity

According to (Sireci, Thissen, & Wainer, 1991) defines a test of validity, namely the extent to which an instrument records or measures what is to be measured. In theory, there are three kinds of instrument validity, namely content validity, construct validity and the last is criteria-based validity. To test the validity, the measuring tool in this study is content validity. Content validity is a concept of measuring validity in which an instrument is considered to have content validity, if it contains questions that are acceptable, the questions are adequate and representative to measure the construct as desired by the researcher (Sireci et al., 1991).

Item Analysis Based on Modern Test Theory

Estimation of items and respondents was carried out using the PROX (normal approximation estimation) procedure. The suitability between the respondent's ability and the item difficulty index (b) will result in accuracy in measurement. Maximum accuracy occurs when P = 0.5. Parameter estimation is done by removing all true and false respondents. The estimation of respondent and item parameters is done simultaneously because it is not yet known. Estimation continues until the respondent and item parameters are constant.

The quality of the item is determined from the suitability of the item with the Rasch model and the item difficulty index. A good item must meet the requirements of item response theory. Items were analyzed for their agreement with the mean-square infit.

Tabel 4. *Item Fit Criteria with the Rasch Model*

Infit Meansquare	*Consideration*
> 1.33	Not suitable
0.77-1.33	Suitable
< 0.77	Not suitable

The next step is to look at the t item outfit value with the following criteria:

Tabel 5. *Item Criteria Accepted and Rejected*

Criteria	*Consideration*
Outfit t $\leq$ 2.00	Accepted
Outfit t > 2.00	Rejected

The third stage is to analyze the Delta or Threshold value (b) with the following criteria.

Table 6. *Categori Threshold (b)*

Threshold	Categori
$b > 2$	Very difficult
$1 < b \leq 2$	Difficult
$-1 \leq B \leq 1$	Moderate
$-1 > b > -2$	Easy
$b < -2$	Very easy

3.Findings

Analysis Results Based on Classical Test Theory

Item Difficulty Level

The item difficulty level of the Quest output file can be seen from the percentage (%) which can be seen in Appendix 23 (Adam & Khoo, 1996). The lowest item difficulty index is 0.13 item 2 and the highest item difficulty index is 0.96 out of item 10. Based on the index, it can be concluded that the most difficult item of the Indonesian achievement test made by the Indonesian Language MGMP is item 2 while the test item is the most easy is item 10. The results of item difficulty level by category in the Indonesian language achievement test are presented in tabular form as follows.

Table 7. *Summary of Item Difficulty Level Results from the Indonesian Achievement Test*

Categori	Item	Total	Percentage
Easy (0.71–1.00)	1,4,5,6,7,9,10,11,12,13,14, 15, 17,18,24, 27, 28,30,31,33,39, 40	22	55
Moderate (0.31- 0.70)	3,16,19,20, 21, 22,23,25, 29,32, 34,35, 36, 37 ,38	15	37.5
Difficult (0.00- 0.30)	2,8,26	3	7.5
	Total	40	100%

Based on table 7. It can be concluded that there are 22 questions that fall into the easy category. The percentage of item difficulty level that is included in the easy category is 22/40 x 100% = 55%. There are 15 items that fall into the medium category. The percentage of item difficulty level that is included in the medium category is 15/40 x 100% = 37.5%. There are 3 items that fall into the difficult category. The percentage of item difficulty level that is included in the difficult category is 3/40 x 100% = 7.5%. Based on the percentage of item difficulty level of each category, it can be concluded that the most dominant item difficulty category in this test is the easy category (55%) and the smallest item difficulty category in this test is the difficult category (7.5 %).

Item Discrimination

Item Discrimination from the Quest file output can be seen from the Point Biser line (Pt-Biserial) which can be seen in Appendix 23 (Adam & Khoo, 1996). The lowest Item Discrimination Index is 0.01, namely item 38 and the highest Item Discrimination is 0.52, namely item 19. The results of the item difficulty level by category on the Indonesian language learning achievement test are presented in tabular form as follows.

Tabel 8. *Summary of Indonesian Language Achievement Test Item Discrimination Results*

Categori	Item	Total	Percentage
Bad (Pt. Biser $\leq$ 0.19)	9,27,31,38	4	10
Acepted (0.20-0.29)	1,2,5,6,8,10,11,15,23,25,30,37	12	30
Good (0.30-0.39)	3,7,13,17,20,33,35,36,39,40	10	25
Very Good (Pt.Biser $\geq$ 0.40)	4,12,14,16,18,19,21,22,24,26,28,29,32,34	14	35
	Total	40	100

Based on table 8, it can be concluded that there are 4 items that fall into the bad category. The percentage of grain differential power that is included in the bad category is 4/40 x 100% = 10%. There are 12 items included in the acceptable category. The percentage of grain differential power that belongs to the accepted category is 12/40 x 100% = 30%. There are 10 items that fall into the good category. The percentage of grain differential power which is included in the good category is 10/40 x 100% = 25%. There are 14 items included in the very good category. The percentage of grain discrepancy that has a very good category is 14/40 x 100% = 35%. Based on the percentage of item differentiating power above, it can be concluded that the most dominant category of item differentiating power in this test is the good category (42.5%) and the smallest item differentiating category in this test is the very good category (12.5%).

Distractor

The distractor is a multiple choice answer that is definitely wrong. Its function is to make students confused or miscalculate when choosing the correct answer among the alternatives provided. A distractor can be an effective distractor or an ineffective distractor. For distractor index analysis, you can see the Quest output file in Appendix 23. An effective distractor index must be more than 0.050 (% 0.050), while an ineffective distractor must be less than 0.050 (% < 0.050) (Suwarto, 2016). Based on the effective distractors and ineffective distractors the Indonesian language achievement test is presented in tabular form as follows.

Table 9. *List of Indonesian Language Achievement Test Distractors*

Item	Ineffective distractor	Effective distractor	Answer key
1	A, B, C	-	D
2	C, D	B	A
3	-	B, C, D	A

4	B	A, D	C
5	C, D	B	A
6	C, D	B	A
7	B, C, D	-	A
8	B	C, D	A
9	B, C, D	-	A
10	A, C, D	-	B
11	C	B, D	A
12	C	A, D	B
13	A, C, D	-	B
14	A, C, D	-	B
15	C	B, D	A
16	-	A, B, C	D
17	A, C, D	-	B
18	B, D	A	C
19	A	B, D	C
20	-	A, C, D	B
21	B, C	A	D
22	-	A, B, C	D
23	-	A, B, C	D
24	A, B	C	D
25	-	A, B, C	D
26	A, C	D	B
27	B, D	C	A
28	D	A, B	C
29	-	A, B, D	C
30	B, C, D	-	A
31	B, C, D	-	A
32	C	A, D	B
33	B	A, C	D
34	-	A, B, D	C
35	-	B, C, D	A
36	-	A, B, D	C
37	-	A, C, D	B
38	-	A, B, D	C
39	A, B	D	C
40	A, B	C	D
Total	56	64	40

Based on table 9 it can be concluded that the Indonesian language achievement test has 56 ineffective distractors and 64 effective distractors. Then the items that have effective distractors are only 12 items. The percentage of ineffective distracters from the test is 56/120 x 100% = 46.70 %. The effective distractor percentage of the test is 64/120 x 100 % = 53.30 %.

Reliability

The reliability of the Indonesian achievement test made by the Indonesian MGMP was 0.990. It can be seen in the Item Estimation Summary from the Quest output file in attachment 23, namely Reliability of Estimate .99 .

Test Validity

Qualitative analysis of grade 7 Indonesian achievement tests at SMP IT MTA Karanganyar includes an analysis of the test grid and the test items themselves. The test grid meets the specification criteria for learning objectives, test grid indicators, and test comprehension. All test items have measured the indicators that should be measured, so that the validity of the Indonesian test is met.

Analysis Results Based on Modern Test Theory

Analysis based on classical test theory has a weakness, namely the characteristics of the items depend on the group of test takers subjected to the items. In the analysis based on the theory of classical statistical test items such as item difficulty index depending on the group of test takers, if the test is carried out by students who are clever then the problem feels easy (higher item difficulty level) and vice versa. conversely if it is done by students who are not good at it then the questions become difficult (the level of difficulty becomes small). Therefore, the characteristics of the test are inconsistent or change depending on the ability of students who take the exam (Hambleton et al., 1991). Therefore, the researcher continues to analyze the characteristics using modern test theory to analyze the characteristics test. Analysis based on modern test theory uses one logistic parameter (1PL) because the Quest program can only analyze one logistic parameter (Adam & Khoo, 1996). Researchers did not analyze the two-parameter logistic model and the three-parameter logistic model.

Table 10. *Categori Threshold (b) Indonesian Achievement Test*

Categori (Criteria)	Item	Total	Percentage
Very difficult ($b > 2$)	2,8,26	3	7.5%
Difficult ($1 < b \leq 2$)	16,20,23,25,34,35,36,37,38,	9	22.5%
Moderate ($-1 \leq b \leq 1$)	3,4,5,11,12,15,18,19,21,22,27,29,32,33	14	35%
Easy ($-1 > b > -2$)	1,6,14,17,24,28,30,31,39,40	10	25%
Very easy ($b < -2$)	7,9,10,13	4	10%
	Total	40	100%

Based on Table 10, the percentage of Indonesian achievement test Threshold = very difficult: difficult: moderate: easy: very easy = 7.5%:22.5%:35%:25%:10%.

If you look at the picture of compatibility with the Racsh Model, the output of the Quest program is as follows.

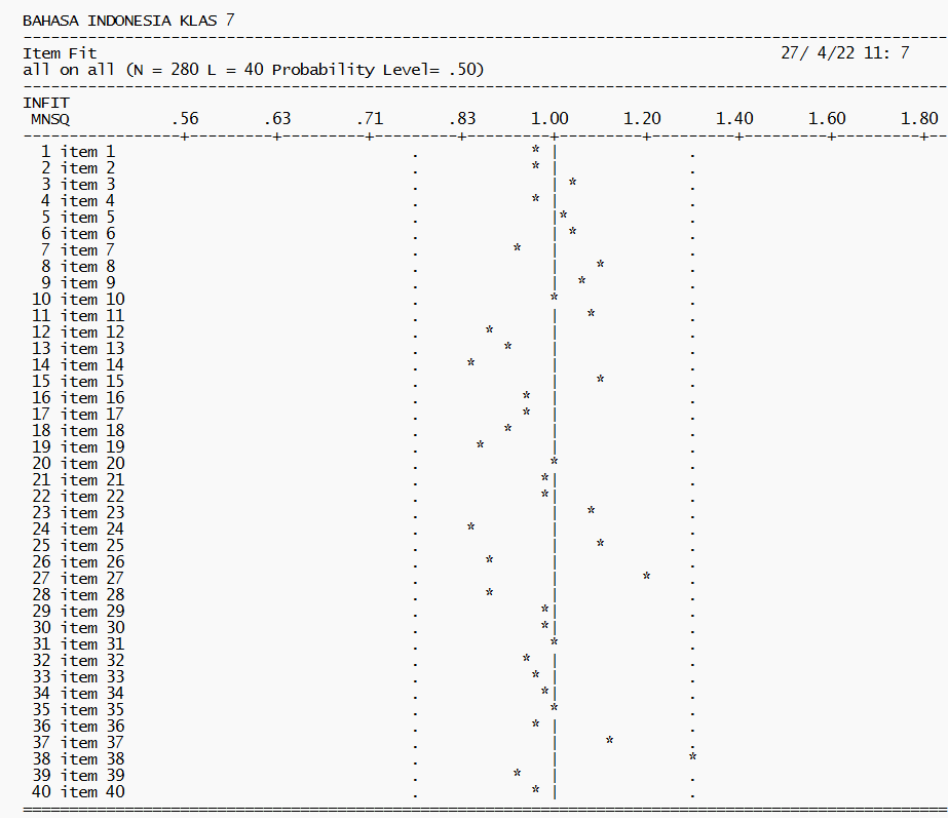

Figure 1. Item fit map for the Indonesian Achievement Test

Based on Figure 1, there are 40 items on the Indonesian language achievement test as an asterisk between two dotted vertical lines (Adam & Khoo, 1996). This shows that all test items fit or match the Rasch Model, all items match 100% (one-parameter logistic model) with the acceptance limit of INFIT MNSQ > 0.77 to <1.30 (Subali & Suyata, 2011).

4.Discussion

This section discusses the findings analyzed by the Quest program with the results of this research related to other studies and other theories. The characteristics of the Indonesian achievement test were identified based on classical test theory: item difficulty level, item discriminatory power, distractibility, reliability. Content validity was analyzed by qualitative analysis with expert judgment. First, item difficulty level consists of 22 easy questions with a percentage of 55%, 15 medium questions with a percentage of 37.5%, 3 difficult questions with a percentage of 7.5%. Based on these results, the difficulty of the item items is more dominant in the easy questions, so that the item does not have proportional item difficulty. An ideal test should consist of 25% easy questions, 50% moderate questions, and 25% difficult questions (Kunandar, 2013); (Suwarto Suwarto, 2016). According to Roid & Haladyna (1982) said that tests that do not have proportional difficulty items, these tests cannot reveal students' actual abilities. This test is also more dominant on easy items where Lund & Winke (2008) H. D. Brown (2003) and Sugianto (2017) emphasize that well-made items should not be too easy or too difficult, the percentage of item difficulty level categories must be balanced, so that tests made by the Indonesian MGMP can identify abilities. or the student's actual Indonesian score. As

stated by Djiwandono (2008) items that can be answered correctly by all test takers, or items that cannot be answered by all test takers are not effective items. According to Suwarto Suwarto (2021), a test consisting of many items is easy to test for students with low achievement, a test that has many items is testing students who have moderate achievement, a test that has many items is difficult to test students with high achievement. Thus, this test is not uniform to assess all students' abilities. This is also supported by Madsen (1983) the percentage of students who answered each item correctly was used by researchers to distinguish between difficult test items and easy test items. The results of the difficulty of these test items can be compared with other studies Masruroh (2014); Haryudin (2015); Putri (2015); Huda & Wahyuni (2019), even though the test conditions are not the same. Previous research found that the level of item difficulty was disproportionate between easy, medium, and difficult items. Item difficulty level can be influenced by cognitive factors (Setiawan, Budi, Sunardi, Gunarhadi, 2021) such as comprehension, coding, transitions, observation, and working memory (Minh et al., 2022). These cognitive factors can affect student learning achievement so that these factors affect the calculation of item difficulty level.

Second, the differential power of the items on this test is good. Based on the output of the Quest program, it shows that there are 4 bad items with a percentage of 10%, 12 items accepted with a percentage of 30%, 10 good items with a percentage of 25%, 14 very good items with a percentage of 35%. These results indicate that 4 bad items must be dropped and 12 acceptable items must be revised (Dichoso & Joy, 2020). The overall results were good because 35% of the items were very good and 25% of the items were dominantly good. This means that most of the items can be stored as a question bank and can be used to measure students' actual Indonesian competence. These items can also distinguish high, medium, and low achievement students. This is in accordance with Suwarto (2021) that the greater the item discriminatory index implies that these items can distinguish between high achieving and low achieving students. This is to detect individual differences among students in the class. The results of this test can be compared with other studies (Boopathiraj & Chellamani, 2013); (Singh et al., 2014); (A. N. S. Saputra, Retnawati, & Yusron, 2021), although the test conditions are not similar. The researchers found good grain discrimination. Meanwhile, different results were obtained from previous studies such as (Sa'adah, 2017); (Toksöz & Ertunç, 2017); (Rehman, Aslam, & Hassan, 2018); (Manalu, 2019); (Maciej Serda et al., 2013) which reported that the grain discriminatory bad, means that the items cannot distinguish between the upper and lower groups.

Third, there are 56 ineffective distractors (46.7%) that need to be revised and 64 effective distractors (53.3%) in this test. The percentage of effective distractors in this study was almost the same as in previous studies, namely Rehman et al (2018). They found 31.07% of cheaters to be effective. In contrast, Toksöz & Ertunç (2017) found more ineffective distractors. Thus, the distractor is ineffective or irrelevant. All of the ineffective distractors support the assertion that all multiple choice items need not be constructed to fulfill the test's objective of providing students with four or more possibilities. In addition, the results of this study can be explained that most of the Indonesian language achievement test items are able to discriminate between high and low achievers which can be assumed that the large item discriminating index can lead to effective distractors (Kheyami, Jaradat, Al-Shibani, & Ali, 2018). They also said that the ideal number of distractors was at least 3 distractors per item.

Furthermore, the reliability of the test is 0.990. This shows that the test items are very reliable. A test with a high level of reliability is classified as a good test item (Sa'adah, 2017). In addition, a good test can be used for next time testing. The results of this study also show that the extent to which Indonesian achievement test measurements remain consistent after repeated tests on subjects and under the same conditions (Suwarto Suwarto, 2021). This reliable test is almost the same as previous studies (Anggreyani, 2009); (Mulianah & Hidayat, 2013); (Pascual, 2016); (A Sugianto, 2020); (Manalu, 2019); (A. N. S. Saputra et al., 2021) although the test conditions are not the same. They found the test reliable. The estimated reliability can be trusted because it is far above the reliability coefficient limit of 0.700. Several factors affect the estimation of reliability, including group homogeneity, time allocation, and test duration. In addition, another factor affecting the estimated reliability is the number of items that are classified as difficult (Crocker, L., and Algina, 1986).

Finally, the Indonesian achievement test is valid because each item on the grid has been represented by an item on the test. These grids meet the specification criteria for learning objectives, indicators, and test comprehension. The results of previous studies such as (Salwa, 2012); (Mulianah & Hidayat, 2013); (Mahirah & Ahmad, 2016); (Manalu, 2019); (Sa'adah, 2017) the summative test they studied was valid. The validity they examine is closely related to the material to be measured in the test. This is supported by Sireci et al (1991) that the test must measure the material contained in the test grid. According to Mahirah & Ahmad (2016) stated that validity is a vital phase in test analysis which assists test makers in determining the suitability of the test with the material. Therefore, test makers such as Indonesian language teachers or Indonesian MGMPs must consider the validity of the tests they create in order to make good tests. Thus, the Indonesian language achievement test can be said to have high validity which can carry out its measurement function, or provide measurement results that are in accordance with the measurement objectives.

5.Conclusion

Characteristics of the Indonesian test: The difficulty level of the test items is 22 items in the easy category with a percentage of 55%, 15 items being moderate with a percentage of 37.5%, and 3 items being difficult with a percentage of 7.5%; The differential power of the items consisted of 4 bad items with a percentage of 10%, 12 items accepted with a percentage of 30%, 10 good items with a percentage of 25%, and 14 very good items with a percentage of 35%; distractors consisted of 64 effective distractors with a percentage of 53.30% and 56 ineffective distractors with a percentage of 46.70%; the test is reliable with a test reliability of 0.990; content validity can be met.
Threshold percentage of Indonesian achievement test = very difficult: difficult: moderate: easy: very easy = 7.5%:22.5%:35%:25%:10%. Based on the Rush Model or one-parameter logistic model, the characteristics of the Indonesian language achievement test of 40 items all match the Racsh Model with a percentage of 100%.

Statements of ethics and conflict of interest

"I, as Corresponding Author, declare and undertake that in the study titled as "The Characteristics of the First Semester Final Test Indonesian Class 7", scientific, ethical and citation rules were followed; Turkish Online Journal of Qualitative Inquiry Journal Editoral Board has no responsibility for all

ethical violations to be encountered, that all responsibility belongs to the author and that this study has not been sent to any other academic publication platform for evaluation."

References

1. Adam, R. J., & Khoo, S.-T. (1996). Acer Quest: The Interactive Test Analysis System. *Australian Council for Educational Research*, pp. 1–96.
2. Ainol Madziah Zubairi, N. L. A. K. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research*, *2*, 1–20.
3. Allen & Yen. (1979). *Introduction to Measurement Theory by Mary J. Allen Wendy M. Yen (z-lib.org)*. Monterey: Brooks/Cole Publishing Company. Retrieved from https://spada.uns.ac.id/pluginfile.php/204338/mod_resource/content/2/Introduction%20to%20Measurement%20Theory%20by%20Mary%20J.%20Allen%20Wendy%20M.%20Yen%20%28z-lib.org%29.pdf
4. Anggreyani, A. (2009). *Dalam Mengevaluasi Butir Soal (Studi Kasus : Soal Ujian Akhir Semester Tingkat Persiapan Bersama Institut Pertanian Bogor Mata Kuliah Fisika Tahun Ajaran 2008 / 2009)*.
5. Bernasela. (2014). An Analysis on English Summative Test Items. *Tanjung Pura University*.
6. Boopathiraj, C., & Chellamani, K. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education. *International Journal of Social Science & Interdisciplinary Research*, *2*(2), 189–193.
7. Brown, H. D. (2003). Language Assessment Principles and Classroom Practice. In *United States of America: Pearson Education*.
8. Brown, J. D. (2001). Point Point--biserial correlation coefficients biserial correlation coefficients. *JLT Testing & Evlution SIG Newsletter*, *5*(3), 13–17.
9. Chauhan, P. R., Ratrhod, P., Chauhan, R., & Rameshbhai, G. (2013). Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot. *Biomirror*, *4*(6), 1–4. Retrieved from www.bmjournal.in
10. Crocker, L., and Algina, J. (1986). Introduction to Classical and Modern Test Theory. In *New York: CBS College Publishing*. Retrieved from http://www.mich.gov/documents/mde/3_Classical_Test_Theory_293437_7.pdf
11. Dichoso, A. A., & Joy, M. R. J. (2020). Test item analyzer using point-biserial correlation and p-values. *International Journal of Scientific & Technology Research*, *9*(4), 2122–2126.
12. Djiwandono, S. (2008). Tes bahasa pegangan bagi pengajar bahasa. *Jakarta: PT Indeks*.
13. Hambleton, R. K., & Rogers, H. J. (1988). Solving criterion-referenced measurement problems with item response models*. *International Journal of Educational Research*, *13*(2), 145–160. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/0883035589900037
14. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
15. Haryudin, A. (2015). Validity and Reliability of English Summative Tests at junior High School in West Bandung. *Jurnal Ilmiah UPT P2M STKIP Siliwangi*, *2*(1), 77–90. https://doi.org/10.22460/p2m.v2i1p77-90.167

16. Hidayati, A. D. (2009). the Analysis of Validity , Reliability , Discrimination Power and Level of Difficulty of First Mid-Term Test in the Case of the Eighth Grade Students of Smp 33 Semarang Faculty of Languages and Arts.
17. Huda, N., & Wahyuni, T. S. (2019). Analisis butir soal IPA Try Out USBN Tahun Ajaran 2018/2019 dalam kaitannya dengan level kognitif. *Madrasah: Jurnal Pendidikan Dan Pembelajaran Dasar*, *12*(1), 29–39.
18. Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple-choice questions at the department of pediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal*, *18*(1), e68.
19. Kolte, V. (2015). Original article Item analysis of Multiple Choice Questions in Physiology examination . *Indian Journal of Basic and Applied Medical Research*, *4*(4), 320–326.
20. Kunandar, K. (2013). Penilaian autentik (Penilaian hasil belajar peserta didik berdasarkan Kurikulum 2013). *Jakarta: Rajawali Pers*.
21. Lababa, J. (2018). Analisis Butir Soal dengan Teori Tes Klasik: Sebuah Pengantar. *Jurnal Pendidikan Islam Iqra'*, *5*, 29–37. https://doi.org/10.30984/jpii.v2i2.538
22. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. L. Erlbaum Associates.
23. Lund, J., & Winke, P. M. (2008). Book review: Brown, H. Douglas (2004). Language assessment: Principles and classroom practices. White Plains, NY: Pearson Education. 324 pp. $48.00 paper. ISBN 0—13—098834—0; Brown, James Dean (2005). Testing in language programs: A comprehensive guide to. *Language Testing*, *25*(2), 273–282. https://doi.org/10.1177/0265532207086784
24. Maciej Serda, Becker, F. G., Cleary, M., Team, R. M., Holtermann, H., The, D., …)2013(.فاطمی, ح. Synteza i aktywność biologiczna nowych analogów tiosemikarbazonowych chelatorów żelaza. *Uniwersytet Śląski*, *7*(1), 343–354. https://doi.org/10.2/JQUERY.MIN.JS
25. Madsen, H. S. (1983). *Techniques in Testing.* ERIC.
26. Mahirah, R., & Ahmad, D. (2016). Designing Multiple Choice Test of Vocabulary for The First Semester Students at English Education Department of Alauddin State Islamic University of Makassar. *ETERNAL (English, Teaching, Learning, and Research Journal)*, *2*(2), 194–208.
27. Manalu, D. (2019). *An Analysis of Students Reading Final Examination by Using Item Analysis Program on Eleventh Grade of SMA Negeri 8 Medan.*
28. Masruroh, H. Z. (2014). An Item Anaalysis on English Summative Test for Second Grade Students of MAN Tulungagung 1 in Academic Year 2013/2014. *A Script: State Islamic Institute Tulungagung*.
29. Minh, A., McLeod, C. B., Reijneveld, S. A., Veldman, K., van Zon, S. K. R., & Bültmann, U. (2022). The role of low educational attainment on the pathway from adolescent internalizing and externalizing problems to early adult labour market disconnection in the Dutch TRAILS cohort. *SSM - Population Health*, 101300. https://doi.org/10.1016/J.SSMPH.2022.101300
30. Mulianah, S., & Hidayat, W. (2013). Pengembangan Tes Berbasis Komputer. *Kuriositas*, *2*(6), 27–43.
31. Mutaqi. (2007). Analisis Butir Soal Terhadap Instrumen Evaluasi Kegiatan Diklat. *Materi Workshop Direktur Diklat Di UDIKLAT PT PLN (PERSERO) Semarang*, 1–10.

32. Pascual, G. R. (2016). Analysis of the english achievement test for ESL learners in Northern Philippines. *International Journal of Advanced Research in Management and Social Sciences*, *5*(12), 1–5.
33. Putri, N. S. (2015). An Analysis of English Semester Test Items based on The Criteria of A Good Test for The First Semester of The First Year of SMK Negeri 1 Gedong Tataan in 2012/2013 Academic Year. *A Script: Lampung University*.
34. Raharja, N. S. (2014). Analisis Butir Soal Ujian Akhir Sekolah Produktif Pemasaran Kelas XII Pemasaran SMK Negeri 9 Semarang. *Economic Education Analysis Journal*, *3*(3), 564–569.
35. Rehman, A., Aslam, A., & Hassan, S. H. (2018). Item analysis of multiple-choice questions. *Pakistan Oral & Dental Journal*, *38*(2), 291–293.
36. Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. Academic Press,.
37. Rusmiana, F. D. (2015). The Test Item Analysis of 1st Semester Final Test of The Accounting Theory for Vocational Education: Case Study of SMK YPKK 1 Sleman Academic Year of 2014/2015. *A Thesis: Yogyakarta State University*.
38. Sa'adah, N. (2017). the Analysis of English Mid-Term Test Items Based on the Criteria of a Good Test At the First Semester of the Eighth Grade Students of MTs . Mathalibul Huda Mlonggo in the Academic Year of 2016 / 2017. *Jurnal Edulingua*, *4*(1), 45–57.
39. Salwa, A. (2012). The validity, reliability, level of difficulty and appropriateness of curriculum of the English test. *Diponegoro University*.
40. Saputra, A. N. S., Retnawati, H., & Yusron, E. (2021). Analysis Difficulties and Characteristics of Item Test of on Biology National Standard School Examination. *6th International Seminar on Science Education (ISSE 2020)*, 8–14. Atlantis Press.
41. Saputra, R. W. (2015). The Comparison Between the Second Mid-Term English Tests for The Seventh Gradersmade by The State and Private School Certified English Teachers (The case of test items analysis of SMP N 1 Semarang and SMP Kesatrian 2 Semarang in the academic year of 2013). *Journal of English Language Teaching*, *4*(1), 1–5. https://doi.org/10.15294/elt.v4i1.7920
42. Setiawan, Budi, Sunardi, Gunarhadi, A. (2021). *Teaching Language Proficiency: The Implementation of Virtual Multimedia-Based Learning for Indonesian Vocational High School*. *48*(11), 289–297. Retrieved from http://jonuns.com/index.php/journal/article/view/865
43. Shomami, A. (2014). An Item Analysis of English Summative Test. *A Script: Syarif Hidayatullah State Islamic University*.
44. Singh, J. P., Kariwal, P., Gupta, S. B., & Shrotriya, V. P. (2014). Original Article Improving Multiple Choice Questions (MCQs) through item analysis : An assessment of the assessment tool. *International Journal of Sciences & Applied Research*, *1*(2), 53–57. Retrieved from www.ijsar.in
45. Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, *28*(3), 237–247.
46. Subali, B., & Suyata, P. (2011). Panduan analisis data pengukuran pendidikan untuk memperoleh bukti empirik kesahihan menggunakan program Quest. *Yogyakarta: Lembaga Penelitian Dan Pengabdian Pada Masyarakat UNY*.
47. Sugianto, A. (2020). *Item Analysis of English Summative Test: EFL Teacher-made Test. Indonesian EFL Research & Practice, 1 (1), 35-54*.
48. Sugianto, Aris. (2017). Validity and reliability of English summative test for senior high school. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, *3*(2), 22–38.
49. Sumadi, S. (2005). Pengembangan alat ukur psikologis. *Yogykarta, Andi Offset*.

50. Suruchi, S., & Rana, S. S. (2012). Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology. *Paripex - Indian Journal Of Research*. https://doi.org/10.15373/22501991/june2014/18
51. Suwarto. (2018). *Statistik Pendidikan*. Yogyakarta: Pustaka Pelajar.
52. Suwarto, S. (2017). Tingkat Kesulitan, Daya Beda, dan Reliabilitas Tes Biologi Kelas 7 Semester Genap. *In Seminar Nasional MIPA 2016*, (November), 312–319. Retrieved from https://conf.unnes.ac.id/index.php/mipa/mipa2016/paper/view/401
53. Suwarto, Suwarto. (2016). Karakteristik Tes Biologi Kelas 7 Semester Gasal. *Jurnal Penelitian Humaniora*, *17*(1), 1. https://doi.org/10.23917/humaniora.v17i1.2346
54. Suwarto, Suwarto. (2017). Pengembangan tes ilmu pengetahuan alam terkomputerisasi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *21*(2), 153–161. Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/13144
55. Suwarto, Suwarto. (2021). The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students Endonezya İ kinci Yar ı y ı l Sekizinci S ı n ı f Ö ğ rencileri İ çin Final S ı nav ı n ı n Özellikleri. *Turkish Online Journal of Qualitative Inquiry (TOJQI)*, *12*(9), 356–370. Retrieved from https://www.tojqi.net/index.php/journal/article/view/5499
56. Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. *Advances in Language and Literary Studies*, *8*(6), 141–146.