e-ISSN 2746-3656
doi: https://doi.org/10.24114/jfi.v2i

A Systematic Literature Review : the Role of Classical Test Theory and Item Response Theory in Item Analysis to Determine the Quality of Mathematics Tests

Hadijah N. I. Siregar, Asmin
Mathematics Education, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Medan (20221), North Sumatera, Indonesia
hadijahsiregar12@gmail.com, asminpanjaitan@gmail.com

Diterima 10 Oktober 2021, disetujui untuk publikasi 28 November 2021

Abstrak. *This research aims to: 1) find out the role of Classical Test Theory and Item Response Theory in item analysis, and 2) find out the role of Classical Test Theory and Item Response theory on the previous research results regarding item analysis to determine the quality of mathematics tests. This research is qualitative research that uses Systematic Literature Review (SLR) as a research method. The research data is collected by documentation. The population of this research is all of the articles from the mathematics journal in the Sinta Ristekbrin database. While the sample is the articles that are obtained from screening. From this research, it can be seen that CTT and IRT play an important role in item analysis, both are commonly used theories and each has its own advantages and disadvantages. The selection of theories to be used in item analysis must consider the advantages and disadvantages of each theory. The indicators used in IRT are validity, reliability, parameter a (distinguishing power), parameter b (level of difficulty), and parameter c (false guess). Based on the results of data analysis, it is known that 8 out of 10 articles use indicators of validity, reliability, level of difficulty, and discriminatory power; 7 out of 10 articles used descriptive quantitative methods; all articles tell about the test form used except article 7 (A7) and article 8 (A8); the number of samples taken affects the implementation of each theory; and only 1 in 10 articles, namely article 10 (A10), provides an overall test quality conclusion.* ***(Jurnal Fibonaci, 02(2): 29 - 48, 2021)***

Kata Kunci: Classical Test Theory (CTT); Item Response Theory; Quality of Mathematics Test; Systematic Literature Review (SLR)

Introduction

The development of science and technology has brought big changes in human life and brought people to global competition (Amalia & Widayati, 2012: 2). Mathematics is also developing and continuing to support the development of science, technology, business dan goverment (Minarni & Napitupulu, 2020 : 3). Mathematics as a field of research taught in formal educational institutions is an important part of efforts to improve the quality of education (Novitasari, 2016: 8).

The success of learning process can be reflected in the results obtained by the participants which can be seen from the results of the evaluation carried out by the teacher (Siswanto, 2006: 60). Changes in students are known from the evaluation (assessment) of the teaching and learning process. The results of tests or evaluations are measuring tools commonly used to determine students' understanding of the material that has been delivered. In addition, from the questions used, it can be seen whether the questions can measure the curriculum objectives that have been set or not, so that the results can be used as a benchmark for the implementation of learning objectives (Hamimi, et al., 2020: 58). Evaluation is very important and must be considered in the learning process. However, many teachers have not been able to choose a good evaluation tool. In general, the teacher-made test in designing questions, did not pay attention or did not analyze the test items so that most of them could not identify the good, mediocre, and bad questions. Apart from analyzing the questions, the tests that are arranged must also meet the requirements or characteristics of good test quality (Supandi & Farikhah, 2016: 72). The test is a measuring tool most often used by teachers to measure student learning outcomes. A new test will be meaningful if it consists of items that test important objectives and represent all the materials being tested (Purwanti, 2014: 82).

Thus, an effort to find out whether the questions made by the teacher are classified as appropriate and good, and provide maximum results in measuring and increasing the level of student understanding, analysis can be carried

out on each item (Sudjana as cited in Rahayu et al., 2014:40). Items analysis that carried out will be able to improve the quality of the questions through elements of validity, reliability, difficulty level, discrimination power, and effectiveness of distractor (Salmina & Adyansyah, 2017:38). Analysis of validity and reliability can be used to determine the quality of the items as a whole, while the analysis of difficulty level, discrimination power and effectiveness of distractor are used to determine the quality of the items. Analysis of difficulty level and discrimination power can be used to measure the quality of objective items and descriptions (Rahayu & Djazari, 2016:86). This is guided by two most frequently used item analysis theories, namely Classical Test Theory (CTT) and Item Response Theory (IRT) as presented by Siri & Freddano (2011:189) and Frey (2017:1).

A more in-depth research is needed regarding the role of Classical Test Theory and Item Response Theory in item analysis. It aims to improve the quality of learning evaluation. Where Systematic Literature Review (SLR) is one method that can be used to examine more deeply about the topic. Therefore, a more in-depth study is needed of the results of previous studies regarding the role of Classical Test Theory and Item Response Theory in item analysis to determine the quality of mathematics tests. Thus, this systematic literature review research can complement the knowledge and address gaps in the literature. This research will present a configuration and conceptual framework for item analysis to determine the quality of mathematics tests related to the role of Classical Test Theory and Item Response Theory for further researches.

Review of Related Theories

Learning

Learning is a process of behavior change caused by experience and training. The experience and training are activities of the teacher as a learner and the activity of students as learners. These behavioral changes can be mental or physical (Sunhaji, 2014:33).

An important factor in learning process is learning objective. With a goal, the teacher has guidelines and principles to be achieved in teaching activities. If the learning objectives are clear and clear, the steps and learning activities will be more focused. Formulated learning objectives should be adjusted to the availability of time, infrastructure and readiness of students. In this connection, all teacher and student activities must be directed at achieving the expected goals (Nata as cited in Pane & Dasopang, 2017:342).

By doing learning we get many benefits. The greatest benefit we get from learning is that we can benefit others. This was expressed by Suyono & Hariyanto in Putria et al. (2020:862), the benefit of learning is obtaining knowledge that is developed through experiences developed through sharing, so that it provides benefits for others.

There are many things related to learning. One of the important things related to learning is about the learning model. For example, based on the distance the learning model can be divided into two, namely face-to-face learning and distance learning (commonly referred to as online learning). Tang & Chaw as cited in Anggrawan (2019:340-341) stated that the face-to-face learning is learning that relies on lecturers presence to teach in class. While online learning is a kind of teaching and learning that uses internet to deliver teaching materials to students (Elyas as cited in Fuadi et al., 2020:194).

Learning Evaluation

Evaluation of education is a process that involves production, application, and instrument analysis of educational measurement. The main function of educational measurement instruments is to offer information on which to base correct decisions when they are made as a means to infer people's capacities (Escudero et al., 2000:2).

According to Riani & Almujab (2020:71), in general, an evaluation of a learning process is carried out to determine the extent to which the objectives have been achieved from the learning that has been done. With the evaluation of a teacher will know a clear picture of the absorption of students they face, the position of students in groups, the strengths and weaknesses of students compared to others, the accuracy or effectiveness of the method used, the difficulty level of the subject matter, the effectiveness and efficiency of the process. learning implemented. The information obtained from these evaluation activities will be useful as feedback material in the learning process. This feedback will be used as an evaluation in the next learning process.

Principles are needed as guides in evaluation activities. Among the evaluation principles are as follows: a) Objective Principle is that evaluation must be carried out objectively. Objective means without influence, because evaluation must be based on real data and must be based on testing that has been carried out. b) Continuous Principle is that evaluation must be carried out continuously. It means that evaluation must be carried out continuously. c) Comprehensive Principle is that evaluation should be carried out comprehensively. This means that the evaluation should be as far as possible regarding all aspects of the student's personality (Subari as cited in Riadi, 2017:4).

The form of evaluation is very crucial to pay attention to because the evaluation results are well and badly influenced by the form of evaluation used (Siswanto, 2006:60). In general, assessment tools (instruments) can be categorized into two forms, namely: 1) Test; and 2) Not a test (non-test). Measurement tools included in the non-test category are: a) Questionnaires; b) Interview; c) Match List (check list); d) Observation; e) Assignment; f) Portfolio; g) Journal; h) Inventory; i) Self-assessment; j) Peer assessment. Whereas a test is a number of questions that must be answered, or questions that must be selected or responded to, tasks that must be done by the person being tested at a certain time. The test is a number of questions that have right or wrong answers, questions that require answers or be given a response to measure a person's level of ability in certain aspects (Wening as cited in Kholis, 2017:307-308).

Item Analysis

Item analysis is a process of examining students' responses to each test item done to measure the quality of the test items. It is a process of checking and analyzing the quality of each item by sorting out the good items from the weak ones and revised them to become better ones (Hartati & Yogi, 2019:60). According to Alpusari (2014:114-115), the use of item analysis is not only limited to improving the items, but there are several things, namely that the item analysis data is useful as a basis: (1) efficient class discussion about test results, (2) for remedial work, (3) for general improvement in classroom learning, and (4) for skills improvement in test construction.

Evaluation through item analysis is very helpful in assessing quality questions so that they are feasible as a measure of student learning success. Analysis of items can be calculated through several elements, namely validity, reliability, difficulty level, discrimination power and distracting function. With the item analysis, good questions and bad questions can be identified as well as which questions can be entered into the question bank, revised or discarded (Salmina & Adyansyah, 2017:38).

Two approaches are widely used for item analysis: (1) the Classical Test Theory that utilizes two main statistics: the item facility index (the percentage of students that correctly answered the item) and the Discrimination index (the point-bacterial relationship between students' performance on individual item and total test score) and (2) the Item Response Theory (IRT) that describes both item statistics and student's ability with the assumption of correlation between the score on a single item and overall test performance. The IRT assumes that there is a correlation between the score gained by a candidate for one (measurable) item/test and their overall ability on the latent trait which underlies test performance (that we want to discover) (Siri & Freddano, 2011:189). This goes along with Frey (2017:1), the most prominent test-theoretical frameworks are Classical Test Theory (CTT) and Item Response Theory (IRT) including the Rasch model.

Classical Test Theory (CTT)

In the early 20th century classical test theory was an emanation. Classical test theory is a ferment of three remarkable achievements from the previous 150 years: The recognition of the existence of measurement error, the concept that error is a random variable, and the concept of correlation and how to calculate it. Then Charles Spearman in 1904 showed how to improve the correlation coefficient to reduce measurement errors and how to obtain reliability to make corrections. Spearman's demonstration marked the beginning of classical test theory. After that, classical test theory was elaborated and refined by Spearman, George Udny Yule, Truman Lee Kelley, and others for the quarter-century or so after 1904. In addition, in 1937 the Kuder-Richardson formula was published. Shortly after that, the next event was the notion of lower bounds (reliability) and the framework for enhancing understanding found in the work of Louis Guttman. The pinnacle of classical test theory was embodied in the systematic treatment received from Melvin Novick (1966) and Lord & Novick (1968) as cited in Traub (1997 : 8).

Classical true-score theory is one of the oldest measurement theories in the world of behavioral measurement. This theory in Indonesian is often referred to as the classical test theory. The classical test theory is a theory that is easy to apply and a model that is quite useful in describing how errors in measurement can affect the score of observations (Sarea & Ruslan, 2019:3).

Classical test theory is a conventional quantitative approach to testing the reliability and validity of a scale based on its items. Classical test theory, also known as *true-score theory*, assumes that each person has a true score, T, that would be obtained if there were no errors in measurement. A person's true score is defined as the expected score over an infinite number of independent administrations of the scale. Scale users never observe a person's true score, only an observed score, X. It is assumed that observed score (X) = true score (T) + some error (E) (Cappelleri et al., 2014 : 649).

The following indicators are generally used in analyzing items based on the classical test theory approach.

a) Validity

Validity reflects the extent to which the accuracy and accuracy of a test instrument serve as a measuring tool for learning outcomes. A test can be said to have validity if the test can measure the object that should be measured and in accordance with certain criteria. A measuring scale or instrument can be said to have high validity if the instrument performs its measuring function, or provides measurement results in accordance with the purpose of the measurement (Amalia & Widayawati, 2012 : 5). According to Sudijono (as cited in Khaerudin, 2015: 216), validity can be seen from two aspects, namely in terms of the test itself as a totality (test validity), and in terms of the items, as an integral part of the test (test item validity). There are several types of test validity used in item analysis according to Andrich & Marais (2019 : 42), namely: content validity, concurrent validity, predictive validity and construct validity. Content validity is known by assessing how relevant the content is by experts based on the operational definition of the trait. In addition, concurrent validity is known by showing the relationship between the results on certain instruments related to the expected way with the results on other relevant instruments. Predictive validity is determined by relating the instrument's results to the future performance of the same nature. The construct validity is known by showing that the results on the instrument are consistent with the expectations of a theoretical understanding of these properties. Meanwhile, according to Khaerudin (2015: 218), the validity of the item can be known through the correlation technique, where an item is said to be valid if the item score has a significant positive correlation with the total score. Then Arikunto (as cited in Alpusari, 2014: 107) grouped the correlation coefficients into several categories, namely: $0.80 < r_{xy} \leq 1.00$ for very high validity, $0.60 < r_{xy} \leq 0.80$ for high validity, $0.40 < r_{xy} \leq 0.60$ for moderate validity, $0.20 < r_{xy} \leq 0.40$ for low validity, and $0.00 \leq r_{xy} \leq 0.20$ for very low validity.

b) Reliability

Reliability is how consistent a person's test score is when repeated measurements are made with the same test or which are considered parallel. In other words, if the test taker gets the same score from two tests of the same or two parallel tests, then the test has perfect reliability ($\rho_{XX'} = 1$). Vice versa, if the test taker obtains a score from a test that is not related at all to the score obtained from another test that is assumed to be parallel ($\rho_{XX'} = 0$), then the two tests are not reliable at all (Allen & Yen as cited in Hayat 2021 : 5). There are several methods that can be used to estimate the reliability of a test. The different methods used affect the interpretation and meaning of reliability which is slightly different. The reliability coefficient which is estimated by giving a test repeatedly to a group of test-takers (test-retest) is defined as the test stability coefficient. While the reliability coefficient obtained from the correlation between subtests or test packages means the test equivalence coefficient (equivalence). On the other hand, the reliability coefficient obtained from giving a test package to a group of people is more accurately interpreted as the coefficient of internal consistency of the test (Crocker & Algina as cited in Hayat 2021 : 5-6).

c) Difficulty Level

The item difficulty level is the ratio of participants who can answer the item correctly from all participants who take the test. In other words, the bigger the index, the easier the item will be because many participants answered correctly and vice versa. The higher the percentage of getting the item right, the easier the item will be. For example, a difficulty index (p-value) of 0.75 means that the item is answered correctly by 75% of the number of test-takers. Then, the difficulty index is grouped into several categories, namely: $p \leq 0.30$ is a difficult item, $0.30 < p \leq 0.70$ is a moderately difficult item, and $p > 0.70$ is an easy item (Matlock-Hetzel as cited in Bichi, 2016 : 28-29).

d) Discrimination Power

Discrimination power is the ability of test items to distinguish high-ability test takers from low-ability test takers. The first step to calculating the discrimination index is to sort the test takers based on their test results. Then take the top 27% test takers and the bottom 27% test takers for further analysis. Thus, the discrimination index is obtained based on the difference between the proportion of the top test takers who answered correctly and the proportion of the lowest test takers who answered correctly. The discrimination index (D) ranges from -1 to +1, where a negative index indicates that most of the lower group answered correctly the test item while a positive index indicates that most of the upper group answered correctly the test item (Courville as cited in Bichi, 2016: 29). Then Ebel & Frisbie (as cited in Bichi, 2016 : 30) grouped the discrimination index into several categories, namely: $D \geq 0.40$ for a very good (satisfactory) test item, $0.30 \leq D \leq 0.39$ for a good test item (no revisions or few revisions), $0.20 \leq D \leq 0.29$ for bad test items and should be revised, and $D \leq 0.19$ for very bad test items (should be deleted or completely revised).

e) Effectiveness of Distractors

According to Anas Sudijono as quoted in Amalia & Widayati (2012 : 10), revealed that the distractor has been able to carry out its function properly if at least 5% of the test takers have chosen it. Distractors that perform well can be reused in future tests. Thus, distractor effectiveness is how well the wrong choice can trick test-takers who do not know the available answer keys. The more test takers who choose distractors, the more distractors can carry out their functions properly. If the test taker ignores all options (doesn't vote) it is called an omit.

The classical test theory has a fundamental limitation, namely the parameter estimation results depend on the characteristics of the examinees (gruop dependent). This implies that the difficulty level of the questions will be low if the test is tested on a group of high-ability test takers and vice versa if the test is tested on participants with low abilities, the level of difficulty of the test will be high. Both results of the estimation of the participant's ability depend on the characteristics of the items (item dependent). This limitation causes the estimation of the participant's ability to be low if the questions given are above their abilities. On the other hand, the estimation of the participant's ability will be high if the questions being tested are below their ability level (Saifuddin as cited in Sarea & Ruslan, 2019:4).

Even so, item analysis using classical test theory is the easiest even though it has several limitations. Some aspects that are considered in the classical test theory are the level of item difficulty, item differentiation, distribution of answer choices, and the reliability of test scores (Safari as cited in Hutabarat, 2009:2).

Overcoming weaknesses in classical test theory, measurement experts develop a model that is not tied to the sample (sample free). This model is hereinafter known as the modern test or item response test. According to item response theory, a person's behavior can be explained by the characteristics of the person concerned to a certain extent (Mardapi as cited in Sarea & Ruslan, 2019:4-5).

Item Response Theory (IRT)

Item response theory (IRT) has finally matured into a powerful and productive alternative to classical test theory and item analysis after years of erratic growth (Bock, 1997). Item Response Theory (IRT) was largely developed in the 1960s to 1980s, as Bock (1997) notes in "Brief Historical Review of Item Response Theory". Thurstone started the foundation of this model in the 1920s. In his paper entitled "A Method of Scaling Psychological and Educational Tests." He provided a technique for placing children's mental development test items on a scale based on age (Binet & Simon, 1905). In addition, Lord & Novick's (1968) book entitled "Statistical Theory of Mental Test Scores" also became the basis for developing the IRT method. They provide a rigorous and statistically unified treatment of classical test theory, particularly the chapters that Birnbaum writes in this book. This explanation is quoted from Bichi & Talib (2018: 143).

Item response theory (IRT) is a collection of measurement models that attempt to explain the connection between observed item responses on a scale and an underlying construct. Specifically, IRT models are mathematical equations describing the association between subjects' levels on a latent variable and the probability of a particular response to an item,using a nonlinear monotonic function (Hays et al. as cited in Cappelleri et al., 2014:654). As in classical test theory, IRT requires that each item be distinct from the others, yet similar and consistent with them in reflecting all important respects of the underlying attribute or construct. Item parameters in IRT are estimated directly using logistic models instead of proportions (difficulty or threshold) and item to scale correlations (discrimination) (Cappelleri et al., 2014:654).

The main purpose of the item response theory being developed is to overcome the classical test theory which is not independent of the group of participants who took the test or the test that was tested. An important part of item response theory is the probability of a test taker's correct answer, item parameters and test taker parameters being linked via a mathematical function or a mathematical formula model. In this formula, the test taker's probability of answering the questions is understood as a logistical function of the different parameters entered into the model (Sarea & Ruslan, 2019:5).

Item response theory or modern test theory was developed on the basis of the following premises: 1) a person's test results can be predicted from their abilities and 2) the relationship between test results and abilities is expressed in a function called the Item Characteristic Curve (Hambleton, Swaminathan, & Rogers, 1991). The function of the item characteristic curve (ICC) shows that the position of test takers with high ability will have a better chance, on the contrary, test takers with a large low ability answer the questions with a high degree of difficulty. This ability is often referred to as potential which is the dominant factor in determining a person's success in learning which

is shown by the results obtained from an exam (Sarea & Ruslan, 2019:5).

In Item Response Theory, item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptotes). The three most commonly used IRT models are; a one-parameter logistic model (1PLM or Rasch model), a two-parameter logistic model (2PLM) and a three-parameter logistic model (3PLM). 1PM has only the difficulty parameter item (b), 2PLM and (b) has a second parameter known as the discrimination parameter (a), which allows the item to differentiate or differentiate examinees with different abilities. 3PLM other than (b) and (a) contains a third parameter, known as the pseudo-probability parameter (c) (Bichi & Thalib, 2018:149).

The relationship with the difficulty level of the items, the invariance nature means that the difficulty index of an item will not change, even if the test taker is smart or less intelligent. This condition does not apply to classical tests so that the invariance is one of the advantages of the item response theory. Modern test theory or item response has more stringent requirements than classical tests, both in terms of assumptions and the sample size required in the analysis (Sarea & Ruslan, 2019:5).

The Item Response Theory measurement model, when compared to the classical model, offers several distinct advantages. These include the following: (a) Question statistics do not depend on the estimated sample (b) Test taker scores do not depend on the difficulty of the test (c) Question analysis accommodates matching items with the examinee's level of knowledge (d) Question analysis does not require parallel testing strictly to assess the reliability (e) The item statistics and the ability of the examinees are both reported on the same scale (Bichi & Thalib, 2018:149).

Sarea & Ruslan (2019:6) say that in item response theory, the mathematical model means that the subject's probability of answering an item correctly depends on the subject's ability and item characteristics. This means that test takers with high abilities will have a greater probability of answering correctly when compared to participants with low abilities. Hambleton et al. (as cited in Sarea & Ruslan, 2019:6) states that there are three assumptions underlying the item response theory, namely unidimensionality, local independence and parameter invariance.

Parameters in Item response theory according to Safaruddin et al. (2012:40), among others: 1). item difficulty level (b), 2). discrimation power items (a), 3). false guess odds (c), 4). participant parameter (θ), and 5). Participants' responses to items are expressed in terms of the probability of answering each step correctly in the item (Pi (θ)).

Item response theory uses the term information to describe test reliability. The information function is very useful for test construction, item selection, measurement precision assessment. Comparison of a number of tests. and determining the weight in the assessment. (Hambleton & Swaminathan as cited in Mardapi, 1998: 29). The amount of information on the test item depends on the discrimination power, the difficulty level, and the pseudo conjecture. The amount of information in principle depends on the level of ability of the test taker. Therefore, to obtain maximum information, the difficulty level of the test must be in accordance with the level of ability that follows the test (Mardapi, 1998:29).

Although IRT itself has several types of models, there are similarities in the indicators used in these models. In fact, the indicators are similar to the CTT. The following are indicators that are commonly used in IRT (Bichi & Talib, 2018 : 147-149).

a) Validity

The meaning of validity and reliability in IRT is different from CTT where IRT focuses on the character of the item. In IRT, validity is the extent to which each examinee and test item ranks well in the ability measured by the test item. Hence, the ability of any test to rank individuals according to their abilities and also to rank items according to their abilities by the level of difficulty.

b) Reliability

In IRT, reliability means the extent to which the score is independent of the group (sample) and of the item. In other words, the characteristics of the test items are independent so that they are not affected by the sample from which they are estimated. In addition, if the same item is given to different groups it gives the same score and rank.

c) The *a* Parameter (Item Discrimination)

The item discrimination parameter 'a' indicates how well the item can discriminate between examinees with different abilities. The value of positive discrimination is if high-ability students have a greater chance of correctly answering the test item and vice versa, low-ability students have a smaller chance of correctly answering the test item. Thus, negative discrimination scores occur if high-ability students have a smaller chance of correctly answering the test item while low-ability students have a higher chance of correctly answering the test item. A test item is good if it has a discrimination value ranging from 0.5 to 2 and the steeper the slope of the item's characteristic curve, the higher the item's discrimination value. The categories of discrimination scores are $a \geq 1.70$ for very good (satisfactory) test items, $1.35 \leq a \leq 1.69$

for good test items (usually without revision), $0.69 \leq a \leq 1.34$ for moderate test items (needs a little revision), $0.35 \leq a \leq 0.64$ for a bad test item (requires many revisions), and $a \leq 0.34$ for a very bad test item (should be removed/replaced).

d) The *b* Parameter (Item Difficulty)

Parameter b refers to the difficulty of the item, is the point where the S-shaped curve has the steepest slope. Only high-ability students can answer difficult test items correctly and low-ability students tend to fail to answer correctly. The category b values are: $-3.00 \leq b \leq -2.00$ for very easy test items, $-2.00 \leq b \leq -1.00$ for easy test items, $-1.00 \leq b \leq 1.00$ for moderately difficult test items, $1.00 \leq b \leq 2.00$ for difficult test items, and $b > 2.00$ for test items are very difficult.

e) The *c* Parameter (Pseudo-Guessing)

The 3PLM (The Three - Parameter Logistics Model) includes a parameter c which is a pseudo-guessing parameter that expresses the probability that a test taker with low ability can get an item correctly and, therefore, has a greater than zero probability of answering an item correctly in a test. The guess parameter c is the lowest value achieved by ICC (Item Characteristic Curves). For example, a student guessing the answer of an item that has four answer choices at random then the probability of guessing correctly is about 0.25.

The IRT model has several technical and practical flaws. The assumptions underlying the use of the IRT model are more stringent than those required in classical test theory. IRT models also tend to be more complex and model outputs more difficult to understand, especially with audiences who are not technically oriented. In addition, the IRT model requires a large sample to obtain accurate and stable parameter estimates, although the Rasch measurement model is useful for small to medium samples. As a result, the choice of model can depend on the available sample, especially in the field testing phase of the certification exam (Bichi & Thalib, 2018:149).

Differences Between CTT and IRT

Based on this explanation, it can be seen that classical test theory has indeed dominated and is widely used in the world of measurement in the last few decades. Almost all concepts of validity and reliability that are known today are developments from the classical test theory. However, classical test theory has several limitations. Therefore, the item response theory (IRT) began to develop, which is a theory developed to correct the limitations of the classical test theory even though the IRT still has limitations.

The following is a comparison between Classical Test Theory and Item Response Theory in terms of various aspects. Table 1 which taken from Muñiz (2010:64) summarizes the differences and similarities between the Classical Test Theory and Item Response Theory.

Table 1 Differences Between Classical Theory and Item Response Theory

Aspects	**Classical Theory**	**Item Response Theory**
Model	Linear	Non-linear
Asssumptions	Weak (easy to fulfil with the data)	Strong (difficult to fulfil with the data)
Measurement invariance	No	Yes
Invariance of test properties	No	Yes
Score scale	Between zero and the maximum test score	Between $-\infty$ and ∞
Emphasis	Test	Item
Item-test relationship	Not specified	Item Characteristic Curve
Item description	Difficulty and discrimination indices	Parameter a, b, c
Measurement errors	Standard error of measurement common to whole sample	Information function (varies according to aptitude level)
Sample size	Can work well with samples of between 200 and 500 participants approx	More than 500 participants recommended, but depends on model

Systematic Literature Review (SLR)

Systematic Review (SR) or usually called as Systematic Literature Review (SLR) is a systematic way to collect, critically evaluate, integrate and present findings from various research studies on research questions or topics of interest. SLRs provide a way to assess the level of quality of existing evidence on questions or topics of interest. SLR provides a broader and more accurate level of understanding than traditional literature reviews (Delgado-Rodríguez and Sillero-Arenas as cited in Nursalam et al., 2020:5).

This is in line with Kitchenham (2004:1), a systematic literature review is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest. Individual studies contributing to a systematic review are called primary studies; a systematic review is a form a secondary study.

The concept of systematic reviewing of research writing got to be powerful within the second half of the 20th century, within the setting of the longstanding, and challenging, issue of how to 'translate' investigate discoveries into the reliable direction for commonsense decision making to decide which approaches, programs, and techniques ought to (and ought to not) be received (Hammersley as cited in Richter et al, 2020:23).

Since the late 1970s, systematic reviews have been utilized within the medical field to supply prove on the viability of practice and treatment. The prove has shown that much of what wellbeing experts do isn't determined from 'what works', but or maybe on what specialists have always done. This finding is not special to the medical calling, additionally happens in other proficient groups including teachers and jail staff. In education, this modern wave of systematic review technique is due in portion to changes in policy towards evidence-based practice: benchmarking and execution markers are being utilized to encourage teachers and educational developers to attain given targets set from national standard benchmarks. In arrange to realize and keep up these targets, teachers and educational developers require data almost which strategies work best in which circumstances, and systematic reviews are one way this data can be given (Perry & Hammond, 2002:32).

The rationale of systematic reviews is that reviews are a frame of research and in this way can be made strides by utilizing suitable and explicit strategies. As the strategies of systematic review have been connected to diverse sorts of research questions, there has been an expanding majority of sorts of the systematic review. In this way, the term 'systematic review' is utilized to allude to a family of research approaches that are a shape of secondary level examination (auxiliary inquire about) that brings together the discoveries of essential research to reply to a research question. Systematic reviews can subsequently be characterized as "a review of existing research utilizing explicit, responsible thorough inquire about methods" (Gough as cited in Richter et al, 2020:4).

In principle, systematic review is a research method that summarizes the results of primary research to present more comprehensive and balanced facts. Meanwhile, meta-analysis is one way to synthesize the results statistically (quantitative technique). Another way to synthesize results is narrative techniques (qualitative techniques) (Siswanto, 2010:329).

A systematic review also has one or more objectives such as: (a) to integrate (compare and contrast) what other people's research has done and said, (b) criticize previous scientific work, (c) to build bridges between related topic areas, and (d) to identify the main problems in a field (Hadi et al., 2020:12).

According to Kitchenham (2004:1-2), there are numerous reasons for undertaking a systematic review. The foremost common reasons are: (1) To outline the existing prove concerning treatment or innovation e.g. to outline the observational prove of the benefits and impediments of spry strategy, (2) To recognize any holes in current inquire about in arrange to propose ranges for further investigation, and (3) To supply a framework/background in arrange to fittingly position modern investigate exercises.

In essence, a systematic review may be a strategy for collecting expansive volumes of information to reach conclusions and suggestions based on the evidence. Systematic reviews are not speedy to conduct and, depending on the degree of the relevant literature, can take months to total. In any case, they are valuable in giving specialists with an evidence base for their hone and will become of expanding significance in case improvements in hone (in anything field) is to be based on soundly looked into prove (Perry & Hammond, 2002:34).

Systematic reviews are widely used by researchers to map areas that are still uncertain, identify research that has been done, and explore new studies that are needed as in the studies above. Systematic reviews can also flag areas of false certainty. These are areas where we think we know more, but in reality there is little evidence to support our beliefs (Petticrew & Roberts as cited in Hadi et al., 2020: 4). So that the review of various studies scattered in various digital libraries is very important in order to find out various kinds of theoretical developments, issues, and research models on certain topics (Hadi et al., 2020:4).

Systematic reviews will be very useful for synthesizing various relevant research results, so that the facts presented to policy makers become

more comprehensive and balanced (Siswanto, 2010:328-329). This is supported by the SLR method carried out systematically by following stages and protocols that allow the article writing process to avoid bias and subjective understanding of the researchers (Nursalam et al., 2020: 5-6).

Article synthesis has many objectives both in terms of historical, conceptual, and methodological understanding: (a) to bring up scientific roots and history on a particular topic, (b) the development of various concepts and debates from various researchers on certain topics, and (c) methods to translate a finding (Hadi et al., 2020:13).

As with individual research methodologies, in principle, systematic review research begins with making a systematic review research protocol and the next stage is carrying out systematic review research (Siswanto, 2010:330). Systematic review applies a strict and transparent methodology in research synthesis to reduce systematic errors (bias) that interfere with the secondary data analysis process (World Health Organization as cited in Hadi et al., 2020:8).

In line with Perry & Hammond (2002:33), the methodology begins with the development of a protocol and search strategy. The protocol outlines the purpose and methodology of the systematic review and is used as a framework to conduct the review procedure. From this, a search strategy is developed and then used and modified to fit the databases specified in the search. Once the potential literature has been identified, the literature is screened against a set of criteria and papers discarded from the review when they do not fit the relevant criteria.

Systematic reviews have strict requirements for search strategy and selecting articles for inclusion in the review, they are effective in synthesizing what the collection of studies are showing in a particular question and can provide evidence of effect that can inform policy and practice (Snyder, 2019:334). The following are 7 stages of the systematic review process according to Cooper namely : 1) Formulating a problem, 2) Looking for literature, 3) Gathering information from articles, 4) Evaluating the quality of research, 5) Analyzing and integrating research results, 6) Interpreting the evidence, and 7) Presentation of results (as cited in Hadi et al., 2020:30-32).

While process stages of systematic review research according to Perry & Hammond (as cited in Siswanto, 2010:330) can be seen in Table 2.

Table 2 Process Stages of Systematic Review Research

No.	Process Stages	Objectives
1	Identification of research questions	To transform research problems into research questions
2	Developing a systematic review research protocol	To provide guidance in conducting systematic reviews
3	Determining the location of the research results database as the search area	To provide a search area limitation to the relevant research results
4	Selection of relevant research results	To collect research results that are relevant to the research question
5	Choosing quality research results	To carry out exclusion and inclusion of research that will be included in a systematic review based on quality
6	Extraction of data from individual studies	To extract data from individual studies to obtain important findings
7	Synthesizing of results by meta-analysis (if it is possible), or narrative methods (if it is not possible)	To synthesize the results using meta-analysis techniques (forest plot) or narrative techniques (meta-synthesis)
8	Presentation of results	To write down the research results in a document that reports on systematic review

Research Methods

Location and Time of Research. The research location is where the researcher conducts research to obtain the desired research data. This research was conducted at internet and Digital Library of

Universitas Negeri Medan for two months from July to September 2021.

Population and Sample. Population is a generalization area consisting of objects or subjects that have certain qualities and characteristics that are determined by the researcher for study and then draw conclusions. While the sample is part of the number and characteristics of the population (Sugiono in Pradana & Reventiary, 2016:4). A large population makes it impossible for researchers to study everything in the population. This is due to limited funds, energy and time. Therefore, the researcher uses a sample from that population provided that the sample from that population must be representative.

The sampling technique used by researcher is purposive sampling (non-probability sampling). Purposive sampling is the sampling technique based on some certain criteria. The population of this research is all articles from mathematics journals on the Sinta Ristekbrin database. After that, the researcher screens the titles and article abstracts. In addition, screening is carried out following the inclusion and exclusion determined by the researcher regarding "item analysis to determine mathematics test quality". So that the appropriate articles will be selected as the sample of this research.

Instruments of Research. Instrument means a tool, so research instruments can be interpreted as tools based on instrument development procedures, theories and research objectives that aim to collect research data. In other words, an instrument is a data collection tool. The instruments used in this study are articles related to "item analysis to determine mathematics test quality" obtained from the Sinta Ristekbrin with screening. In other words, the instruments in this reserach are journal articles related to research questions obtained through a systematic selection procedure.

Design of Research. This research is a qualitative research that uses Systematic Literature Review (SLR) as the research method. According to Delgado-Rodríguez and Sillero-Arenas (as cited in Nursalam et al., 2020:5), SLR is a systematic method for collecting, critically evaluating, integrating and presenting findings from various studies related to research questions or topics of interest. The SLR also provides an assessment of the level of quality of the existing evidence on a question or topic of interest so as to provide a broader and more accurate level of understanding than a traditional literature review.

Research Procedure. The research method used in this research is a systematic literature review. A systematic literature review summarizes the results of primary research to present facts that are more comprehensive and balanced (Siswanto, 2010:329). Like primary research, in principle, a systematic review research methodology begins with making a research protocol and then carrying out the research (Hadi et al., 2020:25). As for the procedure of this research are as follows.

1) Identification of Research Questions

   Before formulating a research problem, first determine the research topic. The research topic is the main core of all the contents of the paper (systematic literature review) to be presented. The topic of this systematic literature review is learning evaluation (mathematics). Furthermore, reading literature activities related to evaluation of learning sourced from scientific articles and books were carried out. So that we get research questions related to item analysis to determine the quality of the math test. The research question is "What is the role of Classical Test Theory and Item Response Theory in determining the quality of the mathematics test in the articles about item analysis used as a systematic review?". The role in question discusses the influence and strengths and weaknesses of the two theories in the analysis of the items, and the situation or the right reasons for choosing the two theories approach. The research questions were obtained based on the PICO (contains of: Population, Intervention, Comparison, Outcome) as in Table 3 below.

   Table 3 Population, Intervention, Comparison, Outcome (PICO)

Population	Butir soal, analisis soal, kualitas soal, karakteristik soal, teori klasik, teori tes klasik, classical test theory, teori respon butir, item response theory
Intervention	Role of Classical Test Theory and Item Response Theory in item analysis
Comparison	Compare with relevant researches
Outcome	The findings regarding the role of Classical Test Theory and Item Response Theory in item

	analysis research articles. For example, in the form of influence and the advantages and disadvantages of each of the two theories and information about the situation or the right reasons for choosing the approach of the two theories.

2) Developing Research Protocol

The research protocol in question is a detailed planning that has been carefully prepared covering several things such as the scope of the research, procedures, criteria for determining the quality and scale of the project. This systematic literature review research protocol follows the inclusion and exclusion criteria as in Table 4 below.

Table 4 Inclusion and Exclusion

PICO	INCLUSION	EXCLUSION
Population	Research on item analysis activities to determine the quality of mathematics tests	Research on item analysis activities to determine the quality of tests other than mathematics subjects
Intervention	Role of Classical Test Theory and Item Response Theory in item analysis	Item analysis approaches except classical test theory and item response theory
Comparison	Relevant researches	Irrelevant researches
Outcome	Results about the effects, strengths and weaknesses, as well as situations or reasons for using Classical Test Theory and Item Response Theory	Results except the effects, strengths and weaknesses, as well as situations or reasons for using Classical Test Theory and Item Response Theory

3) Determining Search Area

At this stage it was determined that https://sinta.ristekbrin.go.id as the location for searching the literature needed to conduct systematic literature review. The Science and Technology Index (SINTA) is a portal that contains the measurement of the performance of Science and Technology including the performance of researchers, writers, authors, journal performance and the performance of science and technology institutions. In addition, SINTA has led to a global (international) indexing portal such as Scopus which has more complete features such as: Citation, Networking, Research and Score. The keywords in the literature search are item items, question analysis, question quality, item characteristics, classical theory, classical test theory, classical test theory, item response theory, and item response theory. Sources of data obtained are focused on scientific articles published in mathematical journals.

4) Selection of Relevant Research Results

The next step is to select research results to obtain relevant research results. Search results for scientific articles on the https://sinta.ristekbrin.go.id database with keywords that match the topic and research questions are checked for duplication of articles, if there is duplication then the article is excluded. Then the researcher conducted a screening based on the title, abstract, and full text whose theme was adjusted to the theme of this systematic literature review. Figure 1 below shows the process of searching and selecting literature.

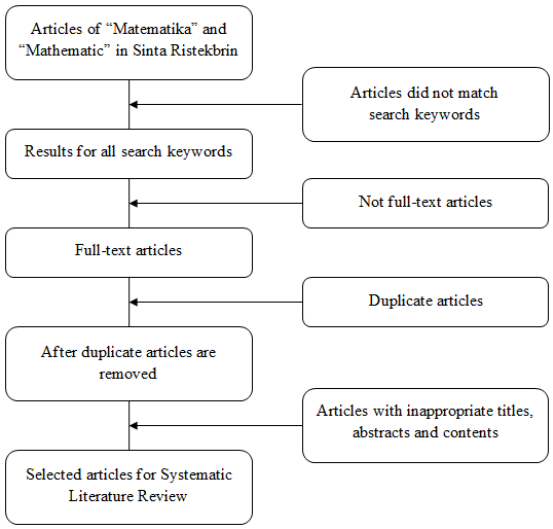

Figure 1 Process of Searching and Selecting Literature

5) Choosing Quality Research Results

After the screening, the researcher then assessed the quality of the articles. This is done to determine the quality of the methodology in each study or article. The methodological assessment is based on theory, design, samples, variables, instruments and data analysis.

6) Extraction of Data

The next step is to extract data then synthesize the various findings found from the previously selected literature. The main purpose of this data synthesis is to analyze and evaluate various research results and literature and to integrate the findings obtained with various disciplines, especially mathematics. This is in line with Hadi et al. (2020: 71), data extraction is the process of retrieving data from data sources for further processing (coding, analysis, and interpretation).

7) Writing and Presentation of Results

After all data extraction and analysis have been carried out, the final stage is the preparation and appearance of the systematic literature review results as a final thesis to complete the undergraduate level. The results of this systematic literature review can later be useful as a source of information in the development of science, especially mathematics education.

Data Analysis. The data analysis technique in this systematic review study uses the Miles and Huberman model, namely data collection, data reduction, data presentation, and drawing conclusion / verification. (1) Data collection, data collection results of research on scientific articles in accordance with research questions. (2) Data reduction, in the form of a summary of research results in scientific articles by selecting the important ones and simplifying them. (3) Data display, is intended to find meaningful patterns and provide the possibility of drawing conclusions or verification and providing action. (4) Drawing conclusions, the final step in data analysis which contains strong evidence that supports the next stage of data collection. Figure 2 below shows the data analysis process according to Miles and Huberman

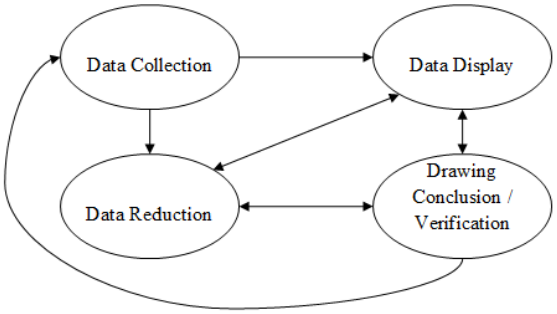

Figure 2 Data Analysis Technique by Miles and Huberman (as cited in Hardani, 2020:163-174)

Result and Discussion

Research Result

This systematic literature review research on the role of classical test theory and item response theory in item analysis to determine the quality of mathematics tests uses the search area on the https://sinta.ristekbrin.go.id page. From this page, 18 articles were obtained that met the requirements for further selection. After screening through titles, abstracts, and contents, 10 articles were selected to be analyzed and synthesized. The process for selecting the articles is described in Figure 3 below.

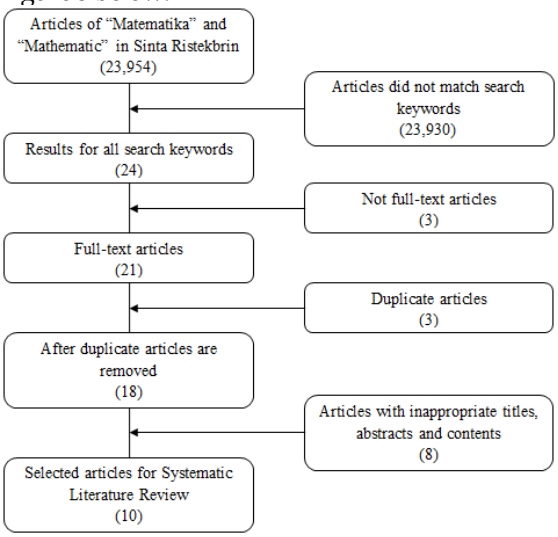

Figure 3 Process of Searching and Selecting Literature

The following is Table 5 which shows the results of the comparison in terms of the roles of CTT and IRT in the 10 scientific articles.

Table 5 Comparison of Articles Based on Role of CTT and IRT

Category / Article	The theory used in research	Indicators used in research	Sampling technique, populations, samples	Software/application used in research	Conclusions are in line with all research objectives
Article 1 (A1)	CTT (not stated directly, known based on indicators)	• Validity • Reliability • Discrimination Power • Difficulty Level	• Technique : purposive sampling • Population : 12 SMP • Sample : 4 SMP (130 siswa)	Microsoft Excel 2007	Not suitable (the results of theoretical and empirical analysis are not mentioned in the conclusion)
Article 2 (A2)	CTT (not stated directly, known based on indicators)	• Curicular Validity • Empirical Validitas • Reliability • Difficulty Level • Discrimination Power • Effectiveness of distractor	• Technique : random sampling • Population : 10.933 students • Sample : 327 students	ANATES	Yes
Article 3 (A3)	CTT	• Material • Construction • Language • Validity • Reliability • Difficulty Level • Discrimination Power • Effectiveness of distractor	• Technique : - (not mentioned) • Population : all of MTs/SMP and MA/SMA in NTB • Sample : 10 schools	Item Test and Analisys (ITEMAN) version 3.00	Yes
Article 4 (A4)	CTT (not stated directly, known based on indicators)	• Reliability • Difficulty Level • Discrimination Power • Effectiveness of distractor	• Technique : purposive sampling • Population : 1 school • Sample : 1 class	Item Test and Analisys (ITEMAN) version 3.50	Not suitable (difficulty index is not mentioned in the conclusion)
Article 5 (A5)	CTT	• Validity • Reliability • Difficulty Level • Discrimination Power • Answer distribution statistics • Measurement error • Score distribution	• Technique :- (not mentioned) • Population : - (not mentioned) • Sample : 43 teachers dan 90 pre-service teacher	Item Test and Analisys (ITEMAN)	Yes
Article 6 (A6)	CTT	• Validity • Reliability • Difficulty Level • Discrimination Power • Answer Distribution	• Technique :- (not mentioned) • Population : SMP in Palopo who took the Mathematics test for the 2014/2015 academic year. • Sample : - (not mentioned)	• MS. Excel • SPSS	Inadequate (modern theory is not mentioned in the goal but is mentioned in the conclusion)
Article 7 (A7)	IRT	• Validity • Reliability • Difficulty Level • Discrimination Power • Pseudo Guessing	• Technique :- (not mentioned) • Population : 23334 college students • Sample : - (not mentioned)	• Microsoft Excel • IBM SPSS 25 • Bilog-MG	Yes
Article 8 (A8)	CTT	• Validity • Reliability • Difficulty Level • Discrimination Power	• Technique :- (not mentioned) • Populasi : 39 people • Sampel : 39 people	-	Yes
Article 9 (A9)	CTT (not stated directly, known based on indicators)	• Validity • Reliability • Difficulty Level • Discrimination Power	• Technique : - (not mentioned) • Population : XII IPS Class in SMA Negari 12 Bandar Lampung for 2014/2015 academic year • Sample : 128 students	• IBM SPSS Statistics 22 • V4	Not suitable (second goal is not answered in the conclusion)
Article 10 (A10)	IRT	• Difficulty Level • Discrimination Power	• Technique :- (not mentioned) • Population : 719 students • Sample : 719 students	Program R	Yes

If we explore further, we will find out to what extent the theory of CTT and IRT is used in each article. This is certainly very useful in knowing the role of CTT and IRT in previous research on item analysis to determine the quality of mathematics tests. The following shows the description of CTT and IRT in the article.

Article 1 (CTT)

In this article, the researcher explains that there are two ways to analyze the items, namely through theoretical/qualitative analysis and empirical/quantitative analysis. Theoretical analysis includes analysis of material, construction, language and level of questions. While the empirical analysis includes validity, reliability, level of difficulty, and discrimination power.

Article 2 (CTT)

In this article, the researcher explained that the ANATES program was believed to be effective and efficient in analyzing items. The data obtained from ANATES are the number of questions analyzed, the number of students, the average correct answer, the standard deviation, the minimum and maximum scores of students, the reliability of the score, the level of difficulty and discrimination. The analysis carried out in this study is curricular validity, empirical validity, test reliability, difficulty level, discrimination power and distractor effectiveness.

Article 3 (CTT)

In this article, the researcher explains that one of the computer programs that can be used to analyze questions based on classical theory is ITEMAN (Item Test and Analysis). Problem analysis can be done qualitatively and quantitatively. Qualitative analysis in question is item analysis based on material, construction and language. While quantitative analysis in the form of analysis of validity, reliability, level of difficulty, discrimination power and functioning of distractors.

Article 4 (CTT)

In this article, the researcher explained that the researcher used the ITEMAN 3.5 application to help analyze the items. The analysis carried out is an analysis of reliability, level of difficulty, discrimination power and effectiveness of distractors.

Article 5 (CTT)

In this article, the researcher explained that to get a good instrument, the instrument must be analyzed qualitatively and quantitatively. Qualitative analysis is a validity analysis that includes material, construction, and language. While quantitative analysis includes analysis of reliability, level of difficulty, discrimination power and functioning of distractors (for multiple choice questions).

Article 6 (CTT)

In this article, the researcher explained that to determine the quality of the test, qualitative (theoretical) and quantitative (empirical) analysis can be used. The qualitative analysis is based on material, language and construction, such as content validity, construct validity, and advance validity. While quantitatively, it is based on classical test theory and item response test theory, such as difficulty, discrimination power, reliability and distractors. Furthermore, the researcher also explained the criteria for determining the quality of items based on the classical test theory which was used as the basis for data analysis to obtain conclusions in this study.

Article 7 (IRT)

In this article, the researcher decided to use item response theory in determining the quality of

the questions. However, the researcher also explained the classical test theory in addition to explaining the item response theory. Classical test theory is widely known and widely used in various fields of science but has a weakness, namely the quality of the test is influenced by the ability of the test taker. Therefore, item response theory was chosen to overcome these weaknesses. Furthermore, the researcher also provides an explanation of the parameters used in item analysis using item response theory. This item response theory is considered more accurate and useful.

Article 8 (CTT)

In this article, the researcher explained that the quality of the test items can be known through their validity, reliability, level of difficulty, and discrimination power. Furthermore, the researcher included the formulas and criteria needed in analyzing the quality of the test questions.

Article 9 (CTT)

In this article, the researcher explained that to determine the quality of the items, it was necessary to carry out quantitative analysis such as validity, reliability, level of difficulty and discrimination power. Furthermore, the researchers included the formulas used in analyzing the items. In addition, the researcher also explained the criteria for determining the quality of test questions such as the correct proportion of easy, medium and difficult questions in a test.

Article 10 (IRT)

In this article, the researcher focused on equalizing the test questions to be studied. That is, the main purpose of this study is to find out whether one test is equivalent to another test. The adjustment in question is based on item response theory. The methods used in the equalization include the moment method (ie: mean, sigma-mean, rigid-sigma mean) and graphic methods (ie: Haebara and Stocking Lord characteristic curves).

Discussion

Role of CTT and IRT in Item Analysis

Classical Test Theory (CTT) and Item Response Theory (IRT) play a very important role in item analysis. We can know this through Siri & Freddano (2011: 189) and Frey (2017: 1) who say that the two types of approaches that are most widely used and best known in item analysis activities are CTT and IRT. In other words, CTT and IRT are very influential in item analysis activities.

The purpose of this study is to see how the role of CTT and IRT in item analysis. That is, this study can show the extent to which the two theories affect the results of item analysis, whether these theories can later show which items are good and which items are bad. If the two theories can show this, the next question is how the two theories differentiate between good items and bad items.

Based on the CTT, either item must meet the following criteria. The criteria that must be met are high validity, high reliability, good discrimination power, and a decent level of difficulty (Alpusari, 2014: 107). In addition, if the test is in the form of multiple-choice, it must have a functioning distractor (Suwarto in Dewi et al., 2019: 16).

The validity of the CTT discusses the accuracy of using a test as a measuring tool for student learning outcomes. A test is said to be valid if the test can measure the object that should be measured according to certain criteria. In other words, a test that achieves the measurement objectives can be said to be a valid test. Vice versa, an invalid test means that the test cannot achieve the measurement objectives (Amalia & Widayawati, 2012: 5). Validity itself has many types, including content validity, concurrent validity, predictive validity, and construct validity (Andrich & Marais, 2019: 42). It is valid in terms of item validity, that is, if the item has a significant positive correlation (Khaerudin, 2015: 218). Based on this explanation, it can be seen that validity can indicate good (valid) items and bad (invalid) items.

Reliability in the CTT discusses how consistent a test result is so that the test can be trusted (reliable) to measure student learning outcomes. Reliability can be done in several methods including test-retest and equivalence coefficient test. Similar and interconnected with validity, items that are said to be reliable are items that have a high-reliability coefficient. Therefore, reliability can show a good item (reliable) and a bad item (unreliable).

Furthermore, there is an indicator of the level of difficulty in the CTT. Items with the highest difficulty level (difficult items) will have a low difficulty index (Matlock-Hetzel as cited in Bichi, 2016 : 28-29). Although this is actually quite debatable because it should be named the level of convenience because in fact, it measures the proportion of students who answer correctly, not

the proportion of students who answer wrongly. A good item is an item that has a moderate level of difficulty (not too difficult and not too easy). But that does not mean that questions that are too easy and questions that are too difficult are not good because they depend on the purpose of the test. No less important is the proportion of easy, medium, and difficult questions that also determine the quality of a test. In other words, a good test is a test that is proportional to easy, medium, and difficult items.

Discrimination power in the CTT discusses the ability of the test in distinguishing high-ability students and low-ability students (Courville as cited in Bichi, 2016: 29). Good items are items that have a significant positive discrimination index. Therefore, discrimination power can show which items are good and which items are bad.

The effectiveness of the distractor in the CTT discusses how many wrong answer choices (detractors) were chosen by students who did not know the correct answer (answer key). A good distractor is a distractor chosen by at least 5% of students (Amalia & Widayati, 2012: 10). Therefore, a good item is an item whose distractors function well.

Based on IRT, a good item must have several criteria. The criteria in question are: high validity, high reliability, good discrimination power, moderate level of difficulty and good probability of guessing correctly (pseudo-guessing).

The meaning of validity and reliability in IRT is different from CTT where IRT focuses on the character of the item, not on the achievement of the test objectives. Validity in IRT is the extent to which each examinee and test item ranks well in the ability measured by the test item. Hence, the ability of any test to sort individuals according to their abilities and also to sort items according to their abilities according to their level of difficulty. So it can be said that a good item is a valid item and a bad item is an invalid item.

Reliability in IRT is the extent to which the score is independent of the group (sample) and item. That is, the characteristics of the test items are independent. Therefore, if the same item is given to different groups it gives the same score and rank. So it can be said that a good item is a reliable item and a bad item is an unreliable item.

Parameter a (item discrimination) in IRT discusses how well the item can distinguish examinees with different abilities. A test item is said to be good if it has a discrimination value ranging from 0.5 to 2 and the steeper the slope of the item's characteristic curve, the higher the item's discrimination value. So that it can be shown good items and bad items.

Parameter b (Item Difficulty) is the point where the S-shaped curve has the steepest slope. Difficult questions can only be answered correctly by high-ability students. A good item is an item with a medium level of difficulty (not too easy and not too difficult).

Parameter c (pseudo-guessing) expresses the probability that a test taker with low ability answers the item correctly and, therefore, has a greater than zero probability of answering the item's question correctly on the test. For example, a student guesses the answer to an item that has four answer choices at random, then the probability of guessing correctly is about 0.25.

Role of CTT and IRT to Determine Quality of Mathematics Tests

We can find out which theory is the most popular between CTT and IRT. This means that we can find out which theory is most often used to analyze the items. If we look back at the articles that we researched, we can briefly see in Table 4.2 and Table 4.3, it is known that CTT is the most popular and frequently used theory of item analysis. In other words, CTT is a theoretical approach that is most often used as an alternative in knowing the quality of a test.

If we pay attention, almost all of the item analysis articles that aim to determine the quality of the mathematics test being studied are 8 out of 10 articles using indicators of validity, reliability, level of difficulty, and discrimination power in analyzing test questions. Article 4 (A4) does not use indicators of validity and article 10 (A10) does not use indicators of validity and reliability. From this, it is known that validity, reliability, level of difficulty, and discrimination power are indicators that are often used in item analysis. Another important indicator that must be taken into account is the effectiveness of distractors for multiple-choice test questions.

A1 uses tests in various forms, namely multiple-choice, short entry, and essay. The indicators used in analyzing the test items include validity, reliability, level of difficulty, and discrimination power. Validity itself includes content validity, construct validity, surface validity, and item validity. The item validity is calculated using the biserial point correlation formula (for multiple-choice questions) and the Pearson product-moment correlation formula (for essay questions). The researcher also lists the test items for the semester exam that do not meet content validity, construct validity, and surface

validity. After that, the researcher explained the reasons why the test items did not meet content validity, construct validity, and surface validity. The researcher quoted Arikunto, Sudjana, Zainul, and Nasoetion's opinion about the proportion of good questions based on the level of difficulty but did not give a final conclusion regarding the results of the analysis. Although indirectly we can know that the questions studied are not proportional (not good).

A2 uses the test in the form of multiple choice. The indicators used in analyzing the test items include validity, reliability, level of difficulty, discrimination power, and distractor effectiveness. The validity itself includes curricular validity and empirical validity. The empirical validity is calculated using the biserial point correlation formula. The researcher quoted Sudjana's opinion about the proportion of good questions based on the level of difficulty but did not give a final conclusion regarding the results of the analysis. Although indirectly we can know that the try-out questions studied are not proportional (not good).

A3 uses a form test that is multiple choice. The indicators used in analyzing the test items include material, construction, language, validity, reliability, level of difficulty, and discrimination power and function of distractors. The quantitative calculations are assisted by the ITEMAN (Item Test and Analysis) program. There is no final conclusion regarding the quality of the questions based on the criteria for good questions.

A4 uses the test in the form of multiple choice. The indicators used in analyzing the test items include reliability, level of difficulty, discrimination power, and effectiveness of distractors. This research was assisted by Item Test and Analysis (ITEMAN) version 3.50. The researcher also gave a final conclusion regarding the quality of the test questions.

A5 uses the test in two forms, namely multiple choice and essay. The purpose of this research is to develop an item analysis module using Item and Test Analysis (ITEMAN). The indicators used in the module are validity, reliability, difficulty level, discrimination power, answer distribution statistics, measurement error, and score distribution.

A6 uses the test in the form of multiple choice. The indicators used in analyzing the test items include validity, reliability, level of difficulty, discrimination power, and distribution of answers. The reliability is calculated using KR21. The researcher cites the opinion of experts in determining the criteria for a good question based on the previously mentioned indicators.

A7 uses the test in the form of multiple choice. This is not stated directly but can be known through the indicators. The indicators used in analyzing the test items include construct validity, reliability, level of difficulty, discrimination power, and pseudo guessing. The researcher cites many opinions from experts including the opinion of Mardapi, Retnawati and Wu regarding the criteria for a good question based on the previously mentioned indicators. In addition, it is also emphasized that the very easy and very difficult items do not always have to be discarded because they can be adapted to the purpose of the test itself.

A8 uses a test in the form of an essay. The indicators used in analyzing the test items include validity, reliability, level of difficulty and discrimination power. The item validity is calculated using the Pearson product moment correlation formula. The researcher also includes examples of valid and invalid test items.

A9 uses the test in the form of multiple choice. The indicators used in analyzing the test items include validity, reliability, level of difficulty, and discrimination power. The validity was calculated using the formula R_x (y-1), Product-Moment, Point Biserial, and IBM SPSS 22 Software. The researcher explained the reasons why the test items did not meet the validity. Reliability is calculated using seven Spearman-Brown, KR-20, KR-21, Rulon, Flanagan, Anova Hoyt, Alpha Cronbach formulas, and two software namely Anates and IBM SPSS 22 software. The researcher also provides an explanation of the advantages and disadvantages of each formula. the. Meanwhile, the level of difficulty and discrimination power is calculated using the Anates program. The researcher also states the proportion of good questions based on the level of difficulty and provides a final conclusion regarding the results of the analysis.

A10 uses the test in the form of multiple choice. The indicators used in analyzing the test items include the level of difficulty and discrimination power. The researcher also lists the difficult test items and their reasons. The purpose of this research is to equalize the USBN test kit (national standard school exam). This equalization is carried out using four methods, namely mean-mean, mean-sigma, Haebara, and Stocking Lord. This equalization is assisted by the R program.

The researcher also explains the advantages and disadvantages of each method used.

As we know that a test consists of several items. So if it is associated with item analysis, we will find out whether an analysis result can provide conclusions about the quality of the test as a whole or only the quality of each item. Based on the articles studied, it is known that only 1 out of 10 articles that is only article 10 (A10) provides an overall conclusion about test quality. While most of them only provide information on the quality of each indicator or item used.

Based on the research method in each article, based on Table 4.1, 7 of the 10 articles used descriptive quantitative methods. The rest used descriptive qualitative research methods. From this, we can see that the descriptive quantitative method is a method that is often used in analyzing items. This is related to the indicators used being directly involved with the numbers and the results need to be described.

According to the form of the test, most of them used multiple-choice tests. Only article 7 (A7), article 8 (A8) and article 10 (A10) did not provide direct information about what form of the test was used in the research. However, we can see that A7 uses a multiple-choice test based on the indicator it uses, namely pseudo-guessing. Likewise with A8 which is known indirectly that the research uses essay through the sample questions included in it. Specifically for A10, we can find out the form of the test through the name of the test, namely USBN (national standard school exam) which is always made by the government in the form of multiple choice. That way, we can know that the form of multiple-choice tests is very commonly used in item analysis activities although item analysis also supports essay tests.

Based on table 4.2, we can see that the sample in the item analysis research to determine the quality of the mathematics test using IRT is more than 500 samples. This is in line with the opinion of Muñiz (2010:64) in table 2.1 which recommends the use of a sample of more than 500 for item analysis using IRT. Thus, it is known that the number of samples affects the implementation of each theory.

Based on these articles, it can be seen that CTT and IRT each have a different perspective and influence on a study. That is, the use of CTT and the use of IRT in item analysis research will each have an impact on determining the quality of a test, in this case especially the mathematics test. So it is necessary to know the advantages and disadvantages of each theory.

CTT is known as the oldest theory in item analysis but is still often used today. The main reason why many teachers or researchers use this theory is because the research steps are simpler and easier to carry out. This is supported by Safari's opinion as cited in Hutabarat (2009: 2) and Sarea & Ruslan (2019: 3). This explanation is also found in article 5 (A5), article 6 (A6) and article 7 (A7).

Although CTT is the easiest and most popular approach in item analysis theory, it has limitations. The most considered limitation is that it depends on the ability of the test taker. In other words, the quality of the test will differ between high-ability test takers and low-ability test takers. This explanation is supported by the opinion of Sudaryono (2011: 722) and Saifuddin as cited in Sarea & Ruslan (2019: 4). Article 6 (A6) and article 7 (A7) also support this explanation.

Although IRT is less popular than CTT, it can overcome the limitations of CTT. The main advantage of IRT is that it does not depend on the ability of the test taker. In other words, IRT is more consistent and more trustworthy. This is supported by the opinion of Sarea & Ruslan (2019: 5) and can also be found in article 7 (A7).

In addition, IRT also has weaknesses. The requirements that must be met in using IRT in item analysis are more stringent and complex. In other words, IRT is a theoretical approach to item analysis that is more difficult to understand and more difficult to implement. This explanation is supported by the opinion of Bichi & Talib (2018: 149).

Advantages and Disadvantages of the Articles

The articles examined in this literature research are diverse and complementary. Each article has its own advantages and disadvantages. The following is a further discussion of the advantages and disadvantages of these articles based on the role of CTT and IRT in their research and how the research was carried out.

What stands out the most here are the many articles that do not convey directly about the theory of the item analysis approach used in the research. This results in a lack of knowledge given to readers regarding the basis or theory in item analysis. However, we can still find out indirectly through the indicators and formulas used in the study as shown in Table 4.2. This is one of the weaknesses of article 1 (A1), article 2 (A2), article 4 (A4) and article 9 (A9). Therefore, this is an

advantage over other articles because it has conveyed directly about what theory was used in his research.

Then there are several articles whose research conclusions are not in accordance with the research objectives. The reason is that there are goals that are not answered/mentioned in the conclusion or there are parts of the conclusion that are different from the objectives. This occurs in article 1 (A1), article 4 (A4), article 6 (A6) and article 9 (A9). In other words, other articles are superior because the research objectives are in sync with the conclusions.

When viewed from the completeness of the media/applications/software used in the research, articles other than article 8 (A8) are superior. With the media/application/software, the research becomes more complete and reduces the human error. In addition, the use of media/applications/software can make research more effective and efficient.

Closing

Conclusion

Classical Test Theory (CTT) and Item Response Theory (IRT) are known for their roles as the most common item analysis approach theories and are often used by teachers/researchers. CTT and IRT are the two leading and most widely used theories in item analysis. Therefore, CTT and IRT play an important role in item analysis. CTT is dependent on test-takers (students) so that the results of item analysis using CTT are inconsistent, very easily influenced by the ability of test-takers (students). The indicators used in the CTT generally are validity, reliability, level of difficulty, discriminatory power, and effectiveness of distractors (specifically multiple-choice tests). From these indicators, we can see the dependence. This is in contrast to the independent IRT. The indicators used in IRT are validity, reliability, parameter a (discrimination power), parameter b (level of difficulty), and parameter c (pseudo guessing). From these indicators, we can see that the results of item analysis using IRT are more consistent.

We know that in the articles on item analysis to determine the quality of the mathematics tests researched: 8 out of 10 articles use indicators of validity, reliability, level of difficulty, and discrimination power; 7 out of 10 articles used descriptive quantitative methods; all articles directly tell about the test form used except article 7 (A7) and article 8 (A8), mostly in the multiple-choice test; the number of samples taken affects the implementation of each theory, where IRT uses a larger sample than the sample used in the CTT; and only 1 in 10 articles provides an overall test quality conclusion, namely article 10 (A10).

CTT is the oldest approach theory but is still popular and used today. Each of these theoretical approaches has a different point of view and influence on the results of research. Both have their advantages and disadvantages. CTT is simpler, easier to understand and to do, but the results of the analysis are very dependent on the ability of the test taker. Meanwhile, IRT does not depend on the ability of the test taker but is more complex and difficult to understand and to do.

Suggestion

To improve the quality of the test, it is very important to pay attention to the theoretical approach used to analyze the items.

In addition, further media / application / software development can facilitate item analysis. Moreover, if teachers are given training in item analysis to improve their knowledge and skills in analyzing items.

References

Afriani, D., Kusno & Ahmad. (2018). Analisis Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran Matematika Kelas VIII SMP K Laster 1 Kabupaten Banyumas Berdasarkan Taksonomi Bloom Dua Dimensi. AlphaMath : Journal of Mathematics Education, 4(1) : 100-114.

Alpusari, M. (2014). Analisis Butir Soal Konsep Dasar IPA Melalui Penggunaan Program Komputer Anates Versi 4.0 For Windows. Jurnal Primary Program Studi Pendidikan Guru Sekolah Dasar Fakultas Keguruan dan Ilmu Pendidikan Universitas Riau, 3(2) : 107-115.

Amalia, A. N. & Widayati, A. (2012). Analisis Butir Soal Tes Kendali Mutu Kelas XII SMA Mata Pelajaran Ekonomi Akuntansi di Kota Yogyakarta. Jurnal Pendidikan Akuntansi Indonesia, X(1) : 2-10.

Andrich, D. & Marais, I. (2019). A Course in Rasch Measurement Theory. Singapore : Springer Nature.

Anggrawan, A. (2019). Analisis Deskriptif Hasil Belajar Pembelajaran Tatap Muka dan Pembelajaran Daring Menurut Gaya Belajar Mahasiswa. Jurnal Matrik, 18(2) : 340-341.

Anggreini, D. & Darmawan, C.A. (2016). Analisis Kualitas Soal Try Out Ujian Nasional dengan Menggunakan Aplikasi Program Anates. Jurnal Pendidikan dan Pembelajaran Matematika (JP2M), 2(1) : 20-34.

Bichi, A.A. (2016). Classical Test Theory: An Introduction to Linear Modeling Approach to Test and Item Analysis.

International Journal for Social Studies, 2(9) : 28-30.

Bichi, A.A. & Thalib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. International Journal of Evaluation and Research in Education (IJERE), 7(2) : 143-149.

Cappelleri, J.C., Lundy, J.J. & Hays, R.D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. Clinical Therapeutics, 36(5) : 649-654.

Courville, T.G. (2004). An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics. Accessed on May 9th 2021 from https://core.ac.uk/download/pdf-/147123147.pdf

Dewi, S. S., Hariastuti, R. M. & Utami, A. U. (2019). Analisis Tingkat Kesukaran dan Daya Pembeda Soal Olimpiade Matematika (OMI) Tingkat SMP Tahun 2018. Jurnal Pendidikan Matematika & Matematika, 3(1) : 16.

Escudero, E. B., Reyna, N. L. & Morales, M. R. (2000). The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination EXHCOBA). Revista Electrónica de Investigación Educativa, 2(1) : 2.

Evendi, E. (2018). Karakteristik Butir Soal dalam Evaluasi Pembelajaran Matematika (Studi Implementasi Praktikum Mahasiswa Jurusan Tadris Matematika di Tingkat MTs/SMP dan MA/SMA se-NTB). Jurnal Riset Teknologi dan Inovasi Pendidikan (JARTIKA), 1(1) : 25-36.

Frey, F. (2017). Test Theory, Classical Test Theory. Accessed on May 20th 2021 from https://www.researchgate.net/publication/31101883

Fuadi, T. M., Musriandi, R. & Suryani, L. (2020). COVID-19 : Penerapan Pembelajaran Daring di Perguruan Tinggi. Jurnal Dedikasi Pendidikan, 4(2) : 194.

Hadi, S., Tjahjono, H.K. & Palupi, M. 2020. Systematic Review: Meta Sintesis Untuk Riset Perilaku Organisasional. Yogyakarta : Vivavictory Abadi.

Hamimi, L., Zamharirah, R. & Rusydy. (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. Mathema Journal, 2020 : 58.

Hardani, et al. 2020. Metode Penelitian Kualitatif & Kuantitatif. Yogyakarta : CV. Pustaka Ilmu.

Hartati, N. & Yogi, H.P.S. (2019). Item Analysis for a Better Quality Test. English Language in Focus (ELIF), 2(1) : 60.

Hayat., B. (2021). Klasika : Program Analisis Item dan Tes dengan Pendekatan Klasik. JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 10(1) : 5-6.

Hodiyanto. (2017). Analisis Butir Soal Pilihan Ganda Matematika Sekolah Menengah Pertama. Jurnal Buana Matematika, 7(2) : 53-60.

Hutabarat, I.M. (2009). Analisis Butir Soal Dengan Teori Tes Klasik (Classical Test Theory) Dan Teori Respons Butir (Item Response Theory) (Studi Kasus: Soal Ujian Olimpiade Sains Provinsi Bidang Informatika 2009). Pythagoras, 5(2) : 2.

Khaerudin. (2015). Kualitas Instrumen Tes Hasil Belajar. Jurnal Madaniyah, 2(IX) : 216-218.

Kholis, R. A. N. (2017). Analisis Tingkat Kesulitan (Difficulty Level) Soal pada Buku Sejarah Kebudayaan Islam Kurikulum 2013. Jurnal Pendidikan Agama Islam, XIV(2) : 307-308.

Kitchenham, B. 2004. Procedures for Performing Systematic Reviews. Australia : Keele University.

Mardapi, D. (1998). Analisis Butir dengan Teori Tes Klasik dan Teori Respons Butir. Jurnal Kependidikan Edisi Khusus Dies, XXVIII : 29.

Minarni, A. & Napitupulu, E. 2020. Developing ICT Integrated Constructivism Based Learning. Medan : Globe Edit.

Muñiz, J. (2010). Test Theories: Classical Theory And Item Response Theory. Papeles del Psicólogo, 31(1) : 64.

Novitasari, D. (2016). Pengaruh Penggunaan Multimedia Interaktif Terhadap Kemampuan Pemahaman Konsep Matematis Siswa. Jurnal Pendidikan Matematika & Matematika, 2(2) : 8.

Nursalam, et al. 2020. Pedoman Penyusunan Literature dan Systematic Review. Surabaya : Fakultas Keperawatan Universitas Airlangga.

Pane, A. & Dasopang, M. D. (2017). Belajar Dan Pembelajaran. Jurnal Kajian Ilmu-ilmu Keislaman, 3(2) : 342.

Perry, A. & Hammond, N. (2002). Systematic reviews: The Experiences of a PhD Student. Psychology Learning and Teaching, 2(1) : 32-34.

Prabowo, A., Sunaryo & Rahmawati, U. (2017). Pengembangan Modul Analisis Butir Soal dengan Menggunakan Item and Test Analysis. AdMathEdu : Jurnal Ilmiah Pendidikan Matematika, Ilmu Matematika dan Matematika Terapan, 7(2) : 99-110.

Pradana, M. & Reventiary, A. (2016). Pengaruh Atribut Produk Terhadap Keputusan Pembelian Sepatu Merek Customade (Studi di Merek Dagang Customade Indonesia). Jurnal Manajemen, 6(1) : 4.

Purwanti, M. (2014). Analisis Butir Soal Ujian Akhir Mata Pelajaran Akuntansi Keuangan Menggunakan Microsoft Office Excel 2010. Jurnal Pendidikan Akuntasi Indonesia, XII(1) : 82.

Putria, H., Maula, L. H. & Uswatun, D.A. (2020). Analisis Proses Pembelajaran Dalam Jaringan (DARING) Masa Pandemi COVID-19 pada

Guru Sekolah Dasar. Jurnal Basicedu, 4(4) : 862.

Rahayu, R. & Djazari, M. (2016). Analisis Kualitas Soal Pra Ujian Nasional Mata Pelajaran Ekonomi Akuntansi. Jurnal Pendidikan Akuntansi Indonesia, XIV(1) : 86.

Riadi, A. (2017). Problematika Sistem Evaluasi Pembelajarn. Ittihad Jurnal Kopertais Wilayah XI Kalimantan, 15(27) : 4.

Riani, D. & Almujab, S. (2020). Analisis Butir Soal dan Kemampuan Siswa dalam Menjawab Soal Ujian Nasional pada Mata Pelajaran Ekonomi. Jurnal Kajian Pendidikan Ekonomi dan Ilmu Ekonomi, IV(1) : 71.

Richter, O.Z., et al. 2020. Systematic Reviews in Educational Research : Methodology, Perspectives and Application. Germany : Springer VS.

Safaruddin, Anisa, M. & Saleh, A.F. (2012). JMSK – Jurnal Matematika, Statistika & Komputasi, 9(1) : 40.

Sainuddin, S. & Ilyas , M. (2016). Karakteristik Butir Tes Matematika Pada Tes Buatan MGMP Matematika Kota Palopo Berdasarkan Teori Klasik. Pedagogy : Jurnal Pendidikan Matematika, 1(1) : 125-146.

Salmina, M. & Adyansyah, F. (2017). Analisis Kualitas Soal Ujian Matematika Semester Genap Kelas XI SMA Inshafuddin Kota Banda Aceh. ISSN 2355-0074, 4(1) : 38.

Santoso, A., Kartianom, K. & Kassymova, G.K. (2019). Kualitas Butir Bank Soal Statistika (Studi Kasus: Instrumen Ujian Akhir Mata Kuliah Statistika Universitas Terbuka). Jurnal Riset Pendidikan Matematika, 6(2) : 165-176.

Sarea, M.S. & Ruslan, R. (2019). KARAKTERISTIK BUTIR SOAL: CLASSICAL TEST THEORY VS ITEM RESPONSE THEORY? Didaktika Jurnal Kependidikan, Fakultas Tarbiyah IAIN Bone, 13(1) : 3-6.

Siri, A. & Freddano, M. (2011). The Use of Item Analysis for the Improvement of Objective Examinations. Procedia – Social and Behavioral Sciences, 29 : 189.

Siswanto. (2006). Penggunaan Tes Essay dalam Evaluasi Pembelajaran. Jurnal Pendidikan Akuntansi Indonesia, V(1) : 60.

Siswanto. (2010). Systematic Review Sebagai Metode Penelitian Untuk Mensintesis Hasil-Hasil Penelitian (Sebuah Pengantar). Buletin Penelitian Sistem Kesehatan, 13(4) : 329-330.

Snyder, H. (2019). Literature Review as A Research Methodology: An Overview and Guidelines. Journal of Business Research, 104 : 334.

Sudaryono. (2011). Implementasi Teori Responsi Butir (Item Response Theory) pada Penilaian Hasil Belajar Akhir di Sekolah. Jurnal Pendidikan dan Kebudayaan, 17(6) : 721-722.

Sunhaji. (2014). Konsep Manajemen Kelas dan Implikasinya dalam Pembelajaran. Jurnal Kependidikan, II(2) : 33.

Supandi & Farikhah, L. (2016). Analisis Butir Soal Matematika Pada Instrumen Uji Coba Materi Segitiga. JIPMat, 1(1) : 71-78.

Supriadi, W. O. S., Rahim, U. & Zamsir. (2018). Kualitas Tes Sumatif Mata Pelajaran Matematika Kelas VIII Semester Genap SMP Negeri 20 Kendari Tahun Pembelajaran 2016/2017. Jurnal Penelitian Pendidikan Matematika, 6(3) : 86.

Susanto, H., Rinaldi, A. & Novalia. (2015). Analisis Validitas Reabilitas Tingkat Kesukaran dan Daya Beda pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. Al-Jabar: Jurnal Pendidikan Matematika, 6(2) : 203-216.

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. Accessed on October 28th 2021 from https://winsteps.com/a/Traub.pdf

Yusron, E., Retnawati, H. & Rafi, I. (2020). Bagaimana Hasil Penyetaraan Paket Tes USBN pada Mata Pelajaran Matematika dengan Teori Respons Butir?. Jurnal Riset Pendidikan Matematika, 7(1) : 1-12.