Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 1 2023, 28 - 38
DOI: *https://doi.org/10.22487/27765660.2023.v3.i1.16417*
eISSN: 2776-5660

ANALYZING THE QUALITY OF MEASUREMENT INSTRUMENTS OF MULTIPLE CHOICE QUESTIONS ON CLASS XI ECONOMICS MATERIAL IN PUBLIC HIGH SCHOOL 3 GORONTALO THROUGH CLASSICAL TEST THEORY AND RASCH MODELS

Luthfiah Yulisharyasti[1*], Ansor Nurdin[2], Nanda Aulia[3], Fhahnul Aiman H. Arfa[4], Fadjryani[5]

[1,2,3,4,5]Statistics Study Program, Faculty of Mathematics and Natural Sciences, Tadulako University, Soekarno-Hatta Street, Palu 94118, Central Sulawesi, Indonesia

***e-mail**: *luthfiahtjang@gmail.com*

ABSTRACT

One form of evaluation of student learning outcomes is the Final Semester Examination. This exam is designed to measure the extent of achievement of educational objectives. A good evaluation must meet several criteria, including good item validity and reliability, a variety of difficulty levels, and the power of differentiation. This study aims to describe the results of a comparative analysis of the quality of measurement instruments in the form of multiple-choice questions using the classical test theory approach and the Rasch model in terms of validity, reliability, difficulty level, and question differentiation. Data were obtained through a website that presents multiple choice exam results of grade XI students at SMA Negeri 3 Gorontalo, consisting of 26 female students and 10 male students. The results showed that in the instrument validity analysis, the Rasch model showed more valid items with a determination category of 0.4 < pt measure corr < 0.8. This means that the Rasch model provides a better analysis compared to the classical test theory analysis. In the reliability analysis, the reliability value of items in the Rasch model is higher but in almost the same category. In analyzing the difficulty level of the instrument, the classical test theory approach shows that the items are in the easy, medium, and difficult categories, so they are still considered capable of measuring students' abilities. However, in the Rasch model, items are only in the very easy, difficult, and extremely difficult categories. In analyzing the power of differentiation, the classical test theory method and the Rasch model have not provided good enough results to identify respondents in several groups based on their level of understanding.

Keywords: *Classical Test Theory, Rasch Model*

Cite: Yulisharyasti, L., Nurdin, A., Aulia, N., Arfa, F. A. H., & Fadjryani. (2023). *Analyzing the Quality of Measurement Instruments of Multiple Choice Questions on Class Xi Economics Material in Public High School 3 Gorontalo Through Classical Test Theory and Rasch Models*. *Parameter: Journal of Statistics*, *3*(1), 28-38, https://doi.org/10.22487/27765660.2023.v3.i1.16417.

[*]Received: Jun 2023; Reviewed: Jun 2023; Published: Jun 2023

INTRODUCTION

One way to assess student learning outcomes is through the End of Semester Examination, which is designed to measure the level of achievement of educational objectives. Consequently, teachers, as the primary agents of the learning system, need to possess not only teaching skills but also the expertise to evaluate the educational process. The success of educational activities can be evaluated based on the assessment results conducted after the educational process activities (Widyaningsih et al., 2018). In addition to evaluating the educational process, educators or teachers are also expected to enhance the assessment tools or instruments used in the educational process, aligning them with the desired educational outcomes. Furthermore, these assessment tools or instruments should be tailored to the measurement methods and information gathering techniques that can effectively indicate the attainment of educational goals.

The methods commonly used in measuring the achievement of educational goals in assessment activities are test and non-test methods. Tests are one way to see the improvement of students' abilities. Tests related to this goal are often called learning achievement tests. Learning achievement tests are tests that are prepared in a planned manner to reveal subject information on materials that have been taught. The learning achievement test is a test used to reveal the level of learning achievement of students. The non-test method is a method of evaluating the learning outcomes of student participants who are tried without "testing" student participants, but by carrying out systematic or known observations by observation, interviews, distributing questionnaires, analyzing scale documents (both behavioral scales and evaluation scales), research problems, and sociometry. Educators' expertise in sorting and controlling suitable methods in assessment activities is a teaching skill that prospective educators must understand (Azwar and Prihartono, 2003).

An effective evaluation tool needs to fulfill certain criteria, such as having strong item validity and reliability, encompassing a range of item difficulty levels, and possessing the capability to differentiate between students who are proficient in answering questions and those who struggle. A valid test is considered good because it accurately measures the intended constructs. The higher the validity and reliability of an assessment instrument, the more valuable the information derived from the research. Validity and reliability are crucial factors in determining the quality of a test (Wahyuningsih and Rosyid, 2015).

Another factor that contributes to the quality of question items is the level of difficulty and the differentiating power of the questions (Wahyuningsih and Rosyid, 2015). The difficulty level of an item is the ratio between the number of students who answer the item correctly and the total number of test participants. A high-quality question falls neither into the category of being too easy nor too difficult. If a question is too easy, it fails to stimulate students to put in extra effort in responding to it, whereas an overly difficult question can discourage students, leading to a lack of motivation to attempt it as it surpasses their capabilities (Iskandar and Rizal, 2018). The differentiating power of an item refers to its ability to distinguish between individuals with high and low levels of proficiency in the aspect being measured within a given group (Bagiyono, 2017). A question exhibits good differentiating power if it effectively distinguishes individuals based on their expertise levels (Sumintono and Widhiarso, 2015). Considering the aforementioned factors, it is important to assess the validity, reliability, difficulty level, and differentiating power of the questions in order to obtain a high-quality test instrument (Perdana, 2018).

Two approaches can be employed to analyze test instruments in the field of learning. The first approach widely used and still prevalent in the field of learning is the classical test theory (CTT). Classical test theory aims to explain measurement error. It utilizes a measurement error model based on the correlation coefficient. The correlation coefficient, originally introduced by Charles Spearman, attempts to elucidate error through two components: true correlation and observed correlation. In the classical test theory approach, the quality of items is primarily determined by their difficulty level and differentiating power. However, item characteristics generated by classical test theory are variable and dependent on the abilities of the test takers (Sarea and Ruslan, 2019).

The second approach involves the utilization of Rasch modeling, which is a more modern approach. Classical test theory exhibits certain weaknesses and limitations in terms of item characteristics. To address these limitations, item response theory has been developed as an alternative theory that overcomes these shortcomings. This theory posits that a test taker's success is solely influenced by their own abilities. The relationship between item success and a person's ability is described by a monotonically increasing function known as the item characteristic function. Rasch modeling presents a different approach to the utilization of scores or raw test information in the context

of learning evaluation. By applying Rasch modeling to raw test information, the aim is to create a measurement scale with equal intervals, which can accurately reflect data regarding the skills of test participants or the quality of questions answered by students. The item analysis conducted with Rasch modeling aims to establish data alignment between item and student characteristics, utilizing the same metric (Sumintono and Widhiarso, 2015).

A previous study conducted by Susdelina et al. (2018) focused on evaluating the quality of instruments used to measure the understanding of quadratic equations using both classical test theory and Rasch models. The findings of the study indicated that the quality of concept understanding measurement instruments, when assessed using the classical test theory approach, demonstrated good validity. However, when the Rasch model was employed, the quality of the measurement instruments was not deemed satisfactory. The reliability of the instruments, as evaluated through both the classical test theory approach (0.536) and the Rasch model (0.71), fell within the moderate category. Regarding the index of difficulty, the classical test theory approach did not yield favorable results, while the Rasch model analysis revealed varying levels of difficulty, including easy, difficult, and very difficult items. The differentiating power of the concept understanding pretest instrument, when assessed using both approaches, was found to be subpar. Previous researchers have yet to explore the comparative analysis of multiple-choice question instruments based on the classical test theory approach and the Rasch model. Thus, this study aimed to conduct a comparative analysis of the instruments' quality in terms of validity, reliability, difficulty level, and distinguishing power using both the classical test theory approach and the Rasch model. The question items examined in this study consisted of multiple-choice questions related to economic material taught in class XI at SMA Negeri 3 Gorontalo. The objective of this item analysis was to assess students' expertise in understanding the taught economic material, contributing to the assessment process.

MATERIALS AND METHODS

Data Sources

The data utilized in this research constitutes secondary data derived from multiple-choice item instruments on economic material at SMA Negeri 3 Gorontalo for the academic year 2021/2022, obtained from the MGMP IPS Indramayu website. The study focused on class XI students at SMA Negeri 3 Gorontalo, encompassing a total of 36 students.

Research Variables

The research variable used is an instrument of multiple choice questions on economic material totaling 33 items. Each question consists of 4 answer choices, namely A, B, C, and D.

Methods

The methods used in this research are Classical Test Theory and Rasch Model.

a. Classical Test Theory

Classical test theory aims to elucidate measurement error by employing a model based on the correlation coefficient. Charles Spearman discovered the correlation coefficient, which comprises two components: true correlation and observed correlation, serving as an explanation for error (Sarea and Ruslan, 2019). Within classical test theory, emphasis is placed on the raw score of an individual test, which reflects a person's ability. Based on this raw score, diverse analyses and interpretations can be derived to suit the requirements of the conducted study (Sumintono and Widhiarso, 2014).

Classical test theory, referred to as classical pure score theory, utilizes a straightforward mathematical framework that establishes a relationship between observed scores (X), true scores (T), and error scores (E) (Wijono and Mardapi, 2016). This theory can be mathematically expressed as follows:

$$X = T + E \quad (1)$$

In this study, the classical test theory approach is employed to assess the characteristics of the test items, including difficulty level, differentiating power, and test reliability. The level of item difficulty is denoted by the symbol Pi, which is a crucial item parameter in test analysis. When the Pi value approaches 0, it signifies that the item is excessively challenging. Conversely, when the Pi value approaches 1, it indicates that the item is overly easy and requires removal or revision. Questions that

are either too difficult or too easy fail to differentiate between the abilities of individual students (Retnawati, 2016). The following formula is utilized to calculate the level of item difficulty.

$$P_i = \frac{\sum B}{N} \quad (2)$$

where:
P_i : the level of success of the i-th question item
$\sum B$: number of test takers who answered the item correctly
N : number of test takers who answered the item

The capacity of an item to distinguish between students with high and low abilities is referred to as differentiating power. The point biserial correlation index can be employed to measure this differentiating power. Several methods are available to determine the magnitude of the discrimination index, including the discrimination index itself, correlation index, and alignment index (Kartowagiran, 2009). The formula utilized to calculate the correlation coefficient of a test item is:

$$r_{pbis} = \left[\frac{\overline{X_i} - \bar{X}}{s_x}\right]\sqrt{\frac{P_i}{1-P_i}} \quad (3)$$

where:
r_{pbis} : the point biserial correlation coefficient
X_i : continuous variable
$\overline{X_i}$: the average X score for participants who answered the item correctly
$\bar{X}$: average of X scores
s_x : standard deviation of the X score
P_i : the proportion of test takers who answered the item correctly

The reliability of a test is generally measured using a numerical coefficient that has a range of -1,00 ≤ ρ ≤ +1,00. A coefficient value that is high indicates a high level of reliability, whereas a coefficient value that is low suggests a low level of reliability. If the reliability is perfect, then the coefficient has a value of +1,00 (Retnawati, 2016). Reliability estimation can be analyzed using the alpha coefficient equation as follows:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum {\sigma_i}^2}{{\sigma_t}^2}\right) \quad (4)$$

where:
α : reliability coefficient
k : number of items
$\sum {\sigma_i}^2$: sum of item variances, $i = 1,2,\ldots,n$
${\sigma_t}^2$: total score variance

The extent to which a question differentiates between students who have grasped the material (competent) and those who have not is indicative of its level of differentiation. Arikunto (2011) outlines the calculation of the differentiation index for objective form tests as follows:

$$D = PA - PB \quad (5)$$

where:
D : Differentiating power sought
PA : Proportion of the upper group who answered correctly
PB : The proportion of the lower group who answered correctly

b. Rasch Model

Classical test theory presents certain weaknesses and limitations in terms of item characteristics. To address these concerns, another theory called item response theory (IRT) has been developed. This

theory suggests that the success of test takers is solely influenced by their individual abilities. The relationship between a person's ability and their performance on each item is described by a monotonically increasing function known as the item characteristic function. In the 1960s, Georg Rasch introduced an analytical model of item response theory called the one-parameter logistic (1PL) model. Later on, Ben Wright popularized this mathematical model. Using raw data in the form of dichotomous responses (true and false) that indicate student abilities, Rasch formulated a model that establishes a connection between students and items (Sumintono and Widhiarso, 2014).

The utilization of the Rasch model in analysis yields fit statistics that offer researchers insights into whether the obtained data adequately represents a pattern in which individuals with higher abilities respond to items in line with their respective difficulty levels. The parameters employed for this purpose include infit and outfit, which are assessed through mean square and standardized values. Infit (also known as inlier-sensitive or information-weighted fit), as described by Sumintono and Widhiarso (2014), refers to the sensitivity of the response pattern to the targeted item in relation to the respondent (person) or vice versa. On the other hand, outfit (outlier-sensitive fit) measures the sensitivity of the response pattern to an item with a particular difficulty level in relation to the respondent or vice versa.

Validity refers to the degree to which research testing instruments accurately measure the intended constructs, enabling accurate conclusions to be drawn from the conducted research sample. Conversely, reliability pertains to the consistency of results produced by a research testing instrument when repeated. Reliability also contributes to the instrument's overall validity. In this study, the Rasch Model approach was employed to assess the validity and reliability of the instruments used. In recent years, the Rasch model, also known as item-response theory (IRT) or latent trait model, has emerged as an alternative framework for understanding measurement and evaluating the quality of instruments or questionnaires. The application of the Rasch model can yield reliable and valid instruments, as it has the capacity to demonstrate high levels of validity and reliability. Consequently, the use of the Rasch model offers a solution to the issue of validity by providing valuable statistics and facilitating a comprehensive examination of instrument validity. Moreover, applying Rasch models in research enables more efficient, reliable, and valid measurements, enhancing the instrument's usability (Yasin et al., 2015).

According to Sumintono and Widhiarso (2014), the Rasch model can serve as a method for restoring data to its natural state. This natural state refers to the fundamental characteristics of continuous quantitative data. Classical measurement theory, which relies on raw data from rating responses, is considered inadequate in reflecting the original properties of continuous quantitative data. Through the utilization of the Rasch model, ordinal responses can be converted into a ratio form with a higher level of precision using probability principles. The analysis employing the Rasch model encompasses five crucial elements: calibration and estimation of item ability, item characteristic curves in the parameter-model, item and instrument information functions, interaction between items and respondents, and item and respondent fit or mismatch. Unlike classical test theory (CTT), the Rasch model analyzes data by examining how well the data aligns with the model, while in CTT, the model is selected based on the available data. Consequently, employing Rasch models in instrument validation furnishes more comprehensive information about the instrument and better fulfills the definition of measurement.

Data Analysis

The methods used in this research are Classical Test Theory Instrument Analysis and Rasch Model. The stages of analysis in this study are:

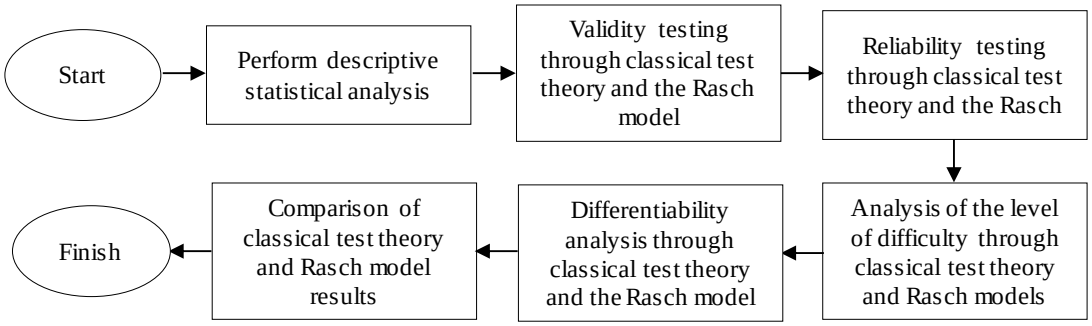

Figure 1. Flowchart of Classical Test Theory Instrument Analysis and Rasch Model

RESULTS AND DISCUSSION

Descriptive Statistical Analysis

Hasan (2001) explains that descriptive statistics, also known as deductive statistics, is a branch of statistics that focuses on the collection and presentation of data in a manner that is easily comprehensible. Descriptive statistics solely concerns itself with describing or presenting information about a given set of data, situation, or phenomenon.

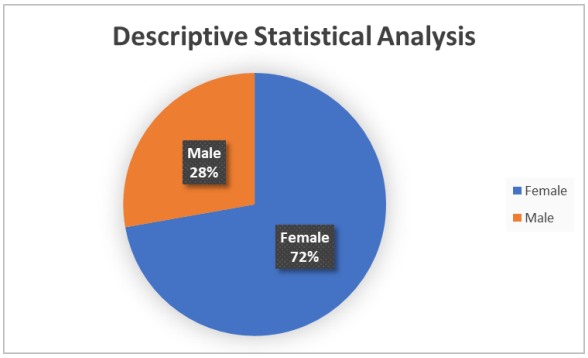

Figure 2. Pie chart Descriptive Statistical Analysis

Respondents in this study were class XI students from SMA Negeri 3 Gorontalo. Based on Figure 1, it can be seen that there are 72% female respondents or around 26 respondents and 28% male respondents or around 10 respondents from a total of 36 respondents.

Validity

The results of the calculation of the validity of the multiple choice test instrument through classical test theory are interpreted with a benchmark if $r_{xy} > r_{tabel}$ on the item, it is said to be valid, otherwise if $r_{xy} \leq r_{tabel}$ on the item, it is said to be invalid. Analysis of the validity of the questions is carried out based on the validity test using the product moment correlation procedure (Syofian et al., 2015). Meanwhile, Sumintono and Widhiarso (2015) propose certain criteria to determine the validity or quality of items in the Rasch model. These criteria are considered satisfied if the items meet the following conditions: (1) The Outfit mean square (MNSQ) value is accepted if $0{,}5 < MNSQ < 1{,}5$; (2) The Z-standard Outfit (ZSTD) value is accepted if $-2{,}0 < ZSTD < 2{,}0$; (3) The Point Measure Correlation (Pt Measure Corr) value is accepted if $0{,}4 < Pt\ Measure\ Corr < 0{,}85$. The comparison of item analysis using Classical Test Theory and the Rasch Model can be seen in table 1.

Table 1. Comparison of Item Validity Analysis Through Classical Test Theory Approach and Rasch Model

No	CTT	Rasch Model	No	CTT	Rasch Model	No	CTT	Rasch Model
1	Valid	Valid	12	Invalid	Valid	23	Invalid	Invalid
2	Invalid	Valid	13	Invalid	Valid	24	Invalid	Invalid
3	Valid	Valid	14	Invalid	Invalid	25	Invalid	Invalid
4	Valid	Valid	15	Valid	Valid	26	Valid	Valid
5	Valid	Valid	16	Invalid	Valid	27	Valid	Valid
6	Invalid	Invalid	17	Invalid	Valid	28	Valid	Valid
7	Valid	Invalid	18	Valid	Valid	29	Invalid	Valid
8	Invalid	Valid	19	Invalid	Valid	30	Invalid	Valid
9	Valid	Valid	20	Invalid	Valid	31	Invalid	Valid
10	Valid	Valid	21	Invalid	Valid	32	Invalid	Valid
11	Invalid	Valid	22	Invalid	Valid	33	Invalid	Valid

Based on the analysis of item instruments using the classical test theory approach with the assistance of IBM SPSS Software version 26, a total of 12 items out of the 33 items examined were found to be valid, as indicated by their item numbers in table 1. Conversely, in the validity analysis using the Rasch model approach with the assistance of R Studio Software, it was found that 27 out of

the 33 items were valid. The Rasch model approach is considered more accurate because it evaluates item validity based on the three aforementioned criteria. Therefore, in addition to its accuracy, the Rasch model analysis also yielded the highest number of valid items compared to the classical test theory analysis. This indicates that the Rasch model offers a superior analysis compared to the classical test theory approach.

Reliability

According to Guilford (1956) in Lestari and Yudhanegara (2017) the reliability coefficient based on classical test theory can be seen using the Cronbach's alpha value. The interpretation of the reliability category can be seen in Table 2.

Table 2. Criteria for Correlation Coefficient of Instrument Reliability of Classical Test Theory

Rate Interval	Categories
0,0 - 0,2	Extremely low
0,21 - 0,4	Low
0,41 - 0,6	Medium
0,61 - 0,8	High
0,81 - 1,0	Extremely high

Meanwhile, according to the views of Sumintono and Widhiarso (2015), the correlation coefficient in the Rasch model is determined by looking at the Item Reliability and Person Reliability values. The interpretation categories can be seen in table 3.

Table 3. Rasch Model Instrument Reliability Correlation Coefficient Criteria

Rate Interval	Categories
< 0,67	Weak
0,67 – 0,80	Simply
0,80 – 0,90	Good
0,91 – 0,94	Very good.
> 0,94	Special

A comparison of item analysis using Classical Test Theory and the Rasch Model can be seen in table 4. The table shows the reliability results consisting of cronbach's alpha, person reliability, and item reliability.

Table 4. Comparison of Reliability Tests Using the Classical Test Theory Approach and the Rasch Model

CTT		**RASCH MODEL**			
Cronbach's Alpha	Categories	*Person reliability*	Categories	*Item Reliability*	Categories
0,639	High	0,47	Weak	0,88	Good

Based on the results of the question reliability analysis as shown in table 4, it can be seen that the reliability through the classical test theory approach shows a high category with a Cronbach's alpha value of 0,639. This means that the consistency of the items answered has a high category. While the Rasch model approach shows a medium category of person reliability and a very high category of item reliability. This means that the consistency of students in answering questions is weak and the consistency of items answered by students is good. From the analysis through these two approaches have different analysis results and different categories. Comparison of the analysis of the two approaches obtained the value of item reliability in the Rasch model provides a higher value but in almost the same category, so it can be said that the item as a measurement instrument is considered an effective and reliable instrument. This means that the instrument should be maintained.

Level of Difficulty

The level of difficulty refers to the likelihood of correctly answering a question at a specific level of ability, typically represented by an index. This difficulty index is commonly presented as a proportion

ranging from 0 - 1. According to Lestari and Yudhanegara (2017) the instrument difficulty index criteria can be seen in table 5.

Table 5. Criteria for Instrument Difficulty Index

Rate Interval	Categories
IK = 0,00	Too difficult
0,00 < IK ≤ 0,30	Difficult
0,30 < IK ≤ 0,70	Medium
0,70 < IK ≤ 1,00	Easy
IK = 1,00	Too Easy

According to Sumintono and Widhiarso (2015), the Rasch model examines the level of item difficulty and categorizes it into four groups based on the measure value obtained in the analysis. These categories are determined as follows: (1) Measure value < -1 = very easy item; (2) Measure value -1 up to 0 = easy item; (3) Measure value 0 up to 1 = difficult item; and (4) Measure value > 1 = very difficult item. The comparison of item analysis using Classical Test Theory and the Rasch Model can be seen in table 6.

Table 6. Comparison of Problem Difficulty Analysis Through Classical Test Theory Approach and Rasch Model

No	CTT	Rasch Model	No	CTT	Rasch Model	No	CTT	Rasch Model
1	Difficult	Too difficult	12	Easy	Too Easy	23	Easy	Too Easy
2	Easy	Too Easy	13	Easy	Too Easy	24	Easy	Too Easy
3	Medium	Difficult	14	Easy	Too Easy	25	Easy	Too Easy
4	Medium	Difficult	15	Difficult	Too difficult	26	Difficult	Too difficult
5	Medium	Difficult	16	Easy	Too Easy	27	Medium	Difficult
6	Easy	Too Easy	17	Easy	Too Easy	28	Difficult	Too difficult
7	Easy	Too Easy	18	Medium	Difficult	29	Easy	Too Easy
8	Easy	Too Easy	19	Easy	Too Easy	30	Easy	Too Easy
9	Difficult	Too difficult	20	Easy	Too Easy	31	Easy	Too Easy
10	Medium	Too difficult	21	Easy	Too Easy	32	Easy	Too Easy
11	Easy	Too Easy	22	Easy	Too Easy	33	Easy	Too Easy

Based on the information presented in table 6, data is gathered to compare the results of item analysis regarding the difficulty level using two different approaches. The comparison between the two approaches reveals that in the classical test theory approach, there are 22 items classified as easy, 6 items classified as medium, and 5 items classified as difficult. On the other hand, the Rasch model shows that there are 22 question items categorized as very easy, 5 question items categorized as difficult, and 6 question items categorized as very difficult. The analysis results from both approaches indicate that the item categories for each item are relatively similar, with only minor variations. The classical test theory approach shows that the question items are still within reasonable limits so that they are still considered capable of measuring students' abilities, while in the Rasch model the question items are only in the very easy, difficult and very difficult categories. The analysis is obtained because the Rasch model performs more accurate analysis compared to classical test theory so that many items from classical test theory have not been detected properly.

The follow-up required after analyzing the level of difficulty is to maintain questions with easy, medium, and difficult levels of difficulty. However, if there are questions that fall into the easy or difficult category, improvements are needed to match the specified indicators. If there are questions that are not feasible or cannot be corrected, they should be deleted and not used again. Furthermore, it needs to be replaced with questions that have better weight and quality so that they can be used again.

Distinguishing Power

According to Arikunto (2011), the differentiating power of an item refers to its capacity to differentiate between students who have a good grasp of the subject matter and those who do not. Table 7 provides the index that indicates the distinguishing power of the instrument.

Table 7. Distinguishing Power Index Criteria

Rate Interval	Categories
0,00 – 0,20	Poor
0,20 – 0,40	Satisfactory
0,40 – 0,70	Good
0,70 - 1,00	Excellent

The results of the calculation of the item differentiation index based on classical test theory and its criteria are presented in table 8 below. The table shows the categories from question discarded to problem accepted/ good.

Table 8. Comparison of Distinguishing Power Through Classical Test Theory and Rasch Model Approaches

Item	Distinguishing Power	Item	Distinguishing Power	Item	Distinguishing Power
1	Satisfactory	12	Poor	23	Poor
2	Poor	13	Poor	24	Poor
3	Good	14	Poor	25	Poor
4	Excellent	15	Satisfactory	26	Good
5	Good	16	Poor	27	Good
6	Poor	17	Poor	28	Satisfactory
7	Poor	18	Satisfactory	29	Poor
8	Poor	19	Poor	30	Poor
9	Good	20	Poor	31	Poor
10	Satisfactory	21	Poor	32	Poor
11	Poor	22	Poor	33	Poor

The results of the analysis of the distinguishing power of questions through the classical test theory approach show that most of the items have poor criteria (discarded questions), namely out of 33 items there are 22 items that have poor criteria. While the statisfacory items are 5 items, the good items are 5 items and as many as 1 item shows that the items are excellent.

In contrast to the Classical Test Theory approach, the Rasch Model analysis aims to differentiate student abilities by examining individual ability levels or employing the respondent separation index. According to Sumintono and Widhiarso (2015), a higher separation value indicates better instrument quality in terms of both respondents and items. This is because a higher separation value enables the identification of distinct groups of respondents and items. The equation used to see the grouping more thoroughly used the equation of stratum separation (H):

$$H = \frac{[(4 \times SEPARATION) + 1]}{3} \quad (6)$$

From this formula, the stratum separation value can be computed to examine variances in student abilities. With an item separation value of 2.77, the H value of 4.03 is rounded to 4. This indicates the presence of four distinct groups of items that can be distinguished. Conversely, with a person separation value of 0.94, the H value of 1.59 is rounded to 2. This suggests that the measurement instrument currently lacks the capability to classify respondents into different groups based on their levels of understanding. Considering the results of the item and respondent separation values, the measurement instrument falls short in adequately identifying groups of items and respondents.

Upon evaluating the two approaches, it is evident that they yield comparable findings concerning item differentiation. The analysis conducted using classical test theory reveals that a significant portion of the questions are categorized as poor, signifying their inability to discriminate between students with

high and low abilities. Similarly, the analysis employing the Rasch model also concludes that the items within the measurement instrument lack the proficiency to classify respondents into distinct groups based on their level of understanding. Thus, both approaches indicate that the measurement instrument's ability to differentiate respondents based on their abilities remains inadequate.

This study shows differences in results in terms of validity, reliability, difficulty level, and distinguishing power with previous research conducted by Susdelian et al (2018), where the results obtained in previous studies were in terms of the validity of concept understanding measurement instruments through the classical test theory approach has good quality in terms of validity while through the Rasch model does not have good quality. While in this study the Rasch model shows more valid items, this means that the Rasch model provides better analysis compared to classical test theory analysis. In instrument reliability through the classical test theory approach (0,536) and the Rasch model (0,71) are included in the moderate category, while in this study the reliability value of the items in the Rasch model provides a higher value but in almost the same category including in the good category. Based on the index of difficulty through the classical test theory approach does not have good quality, while the results of analysis through the Rasch model show varying levels of difficulty, namely easy, difficult and very difficult. While in this study the level of instrument difficulty, it can be seen that the classical test theory approach shows that the items are in the easy, medium, and difficult categories so that they are considered still able to measure student abilities, while in the Rasch model the items are only in the categories of very easy, difficult, and very difficult. As for the analysis of the distinguishing power of the concept understanding pretest instrument through the analysis of the two approaches is not good, while in this study that through the classical test theory method and the Rasch measurement model is still not good enough to identify respondents into several groups based on their level of understanding.

CONCLUSION

Based on the analyzed data comparing Classical Test Theory and Rasch Model, it can be concluded that in the descriptive analysis, approximately 72% of the respondents (around 26 out of 36) are female, while 28% (around 10 out of 36) are male. Regarding the validity test, the Rasch model analysis is considered more accurate since items are deemed valid when they meet the previously mentioned three criteria. Consequently, in addition to its accuracy, the Rasch model indicates a higher number of valid items, implying that it provides a superior analysis compared to the classical test theory. In terms of reliability test, the item reliability value obtained from the Rasch model is higher but falls within the same category. Thus, the measurement instrument is considered effective and reliable, suggesting that it should be maintained. Concerning the difficulty level analysis, the classical test theory approach reveals that the question items remain within acceptable limits, indicating their ability to measure students' abilities. On the other hand, the Rasch model categorizes question items as very easy, difficult, or very difficult. This analysis is obtained because the Rasch model offers a more accurate analysis compared to classical test theory, resulting in the proper identification of many items that were previously undetected. Lastly, in the differentiating power analysis, both the classical test theory and Rasch model methods indicate that most of the questions are classified as poor, indicating their inability to distinguish between students with high and low abilities.

SUGGESTION

Based on the research that has been done, it is hoped that it can be a reference for readers in making a question item instrument as a measurement of student abilities in this case in the form of multiple choice. Thus, the results of measuring student abilities that are used in the future are truly able to measure student abilities properly.

REFERENCES

Arikunto, S. (2011). *Penilaian dan penelitian bidang bimbingan dan konseling*. Yogyakarta: Aditya Media.

Azwar, A., & Prihartono, J. (2003). Metodologi penelitian kedokteran dan kesehatan masyarakat. *Batam: Binarupa Akara*.

Bagiyono, B. (2017). Analisis Tingkat Kesukaran dan Daya Pembeda Butir Soal Ujian Pelatihan Radiografi Tingkat 1. *Widyanuklida*, *16*(1).

Grossman, R. I., Lenkinski, R. E., Ramer, K. N., Gonzalez-Scarano, F., & Cohen, J. A. (1992). MR proton spectroscopy in multiple sclerosis. *American journal of neuroradiology*, *13*(6), 1535-1543.

Hasan, Iqbal. (2001). *Pokok-Pokok Materi Statistik 1 (Statistik Deskriptif)*. Jakarta : PT Bumi Aksara.

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12-23.

Kartowagiran, B. (2009, November). *Penyusunan Instrumen Kinerja* SMK-SBI. In *Makalah dalam Workshop Evaluasi Kinerja SMK-SBI P4TK Matematika*.

Lestari, K. E., & Yudhanegara, M. R. (2017). Analisis kemampuan representasi matematis mahasiswa pada mata kuliah geometri transformasi berdasarkan latar belakang pendidikan menengah. *Jurnal Matematika Integratif*, *13*(1), 28-33.

Perdana, S. A. (2018). Analisis kualitas instrumen pengukuran pemahaman konsep persamaan kuadrat melalui teori tes klasik dan rasch model. *Jurnal Kiprah*, *6*(1), 41-48.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometrian)*. Parama publishing.

Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory Vs Item Response Theory?. *Didaktika: Jurnal Kependidikan*, *13*(1), 1-16.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi)*. Trim Komunikata Publishing House.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.

Susdelina, P. SA, & Febrian.(2018). Analisis kualitas instrumen pengukuran pemahaman konsep persamaan kuadrat melalui teori tes klasik dan rasch model. *Jurnal Kiprah*, *6*(1), 41-48.

Syofian, S., Setiyaningsih, T., & Syamsiah, N. (2015). Otomatisasi metode penelitian skala likert berbasis web. *Prosiding Semnastek.*

Wahyuningsih, T., & Rosyid, R. (2015). Faktor yang Mempengaruhi Kesulitan Belajar Siswa pada Mata Diklat Siklus Akuntansi Kelas XI di Smk. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK)*, *4*(9).

Wijono, S., & Mardapi, D. (2016). Model evaluasi ujian nasional kompetensi keahlian teknik pemesinan SMK. *Jurnal Penelitian dan Evaluasi Pendidikan*, 20(2), 234-243.

Yasin, M., Rajendran, J. J., Sinanoglu, O., & Karri, R. (2015). On improving the security of logic locking. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 35(9), 1411-1424.

Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of e-learning based-STEM on quantum physics subject to student HOTS ability.