JERE 11 (1) 2022 103-113

Journal of Educational Research and Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

Validity and Reliability Analysis Using the Rasch Model to Measure the Quality of Mathematics Test Items of Vocational High Schools

Nur Romdlon Maslahul Adi[1]✉, Hidar Amaruddin[2], Habib Maulana Maslahul Adi[3], Isna Laili Qurroti A'yun[4]

[1]Universitas Islam Negeri Sunan Ampel, Surabaya, Indonesia
[2]Universitas Nahdlatul Ulama Yogyakarta, Indonesia
[3]Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, Indonesia
[4]Universitas Diponegoro, Semarang, Indonesia

Article Info

History Articles
Received:
10 May 2022
Accepted:
12 June 2022
Published:
30 August 2022

Keywords:
Item response theory, rasch model, winsteps

Abstract

Teachers are expected to be able to evaluate the learning process and also be able to develop evaluation tools or instruments used in the learning process. Test items are considered appropriate if they are valid and able to measure the level of suitability and reliability or consistency. The purpose of this study was to analyze the quality of the test items and the mathematical ability of the XII grade students of the State Vocational School in one of State Vocational School in Semarang District. This research is a quantitative descriptive study that describes the quality of the mathematics test items made by the teacher and the students' abilities in mathematics. The subjects in this study were 126 student response patterns to the test instrument in the form of multiple choice questions totaling 20 items for class XII students. Rasch model was used to analyze this research with the help of Winsteps 3.73 software. A total of 20 items were tested, there was only one item, namely item number 7, which did not fit, which means it does not function normally to measure students' abilities. The ability of students who are relatively weak with Person Reliability of 0.48. The results showed that the quality of the test items was applicable. This is indicated by the Item Reliability value of 0.95 which means that the items are very good for evaluating students' abilities.

✉Correspondence Address :
Universitas Islam Negeri Sunan Ampel Surabaya
Jl. Ahmad Yani No.117, Jemur Wonosari, Kec. Wonocolo, Kota Surabaya, Jawa Timur 60237
E-mail : nur.romdlon.maslahul.adi@uinsby.ac.id

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Evaluation is a very important requirement that critically examines the learning process. Evaluation is used to determine whether the educational objectives implemented have been achieved. Therefore, evaluation is a measure of whether teachers are able to teach well and have the ability to evaluate the learning process.

A good evaluation process is certainly carried out with good instruments. Educators or teachers are not only expected to be able to evaluate the learning process, they must also be able to develop evaluation tools or instruments used in the learning process according to the type of expected learning outcomes. more than that, evaluation tools or instruments must also be adapted to techniques in measuring and obtaining data which can then serve as an indicator of the achievement of learning objectives (Erfan et al., 2020).

An assessment requires a good and appropriate tool in measuring students' abilities. One of the tools used is a test. In order for the tests to be carried out to determine the actual abilities of students, several criteria must be met. A good test instrument is one that meets several criteria, including good item validity, good item reliability, varying levels of item difficulty, and has different power of questions that are able to distinguish smart students and are able to answer questions with students. who are unable to answer the questions (Fauziana et al., 2021).

The item is said to be valid if the item measures what it is supposed to measure. That is, if the desired learning outcomes include changes in knowledge, skills, and attitudes, then the questions made must also include these three things (Sumintono & Widhiarso, 2015). Reliability testing is to determine the consistency of the measuring instrument, whether the measuring instrument used is reliable and remains consistent if the measurement is repeated (Hazlita et al., 2014).

However, in fact, the teachers in the process of preparing the test instrument did not conduct an analysis of the quality of the items in the form of an analysis of validity and reliability. The preparation of the test instrument is carried out by the teacher based on the grid that has been made. This is also what was done at the State Vocational School where this research was conducted. The quality of the test instruments designed to test students' final semester abilities has not been analyzed for quality, so it is necessary to analyze the quality of the instruments in the form of validity and reliability analysis. According to Fauziana et al., (2021), the test instrument that has not been analyzed for the quality of the item results in a quasi-assessment which has an impact on the unmeasured ability of the actual student.

Analysis of the quality of test instruments can be done with two approaches, namely classical test theory or Classical Test Theory (CTT) and modern test theory or commonly called Item Response Theory (IRT) which usually uses Rasch modeling. Classical Test Theory (CTT) is a basic method that is generally used in item analysis (Fernanda & Hidayah, 2020).

The current test instrument quality analysis approach that is still widely used is the classical test theory approach or CTT. CTT is used to predict test results by considering several parameters as students' abilities and the level of difficulty of the item (Sumintono, 2018). CTT in its development has received a lot of criticism because it has limitations such as the score obtained depends on the sample and the central score. In addition, the measurement of reliability on the CTT using Cronbach's Alpha and validity is based on the correlation of the scale score with other measurements that are not necessarily reliable or valid (Van Zile-Tamsen, 2017). In the analysis using CTT, students' abilities are only seen from the total score without paying

attention to the correlation between the ability of test takers and the characteristics of the items (Pratama, 2020).

Many researchers found the weakness of the CTT, which led to an alternative item response theory (IRT) analysis. IRT does not depend on the sample of certain question items and the abilities of the people involved in the test (Sumintono, 2018). There are three types of logistic parameter models in IRT: the first, the 1PL model which involves one parameter in the form of difficulty level; the second, the 2PL model which involves two parameters, namely item difficulty level and item discrimination power; and the third, the 3PL model which involves three parameters, namely item difficulty level, item discrimination power, and pseudo guess (Sumintono & Widhiarso, 2015).

The 1PL model was developed by Georg Rasch, a mathematician from Denmark, which was later called the Rasch Model (Sumintono, 2018). Rasch (1960) in Bond & Fox (2015) explains that a person is said to have more ability than others if he has a greater chance of answering one item correctly. Items with a high level of difficulty make the individual's opportunity to answer smaller. The Rasch model only uses one parameter, namely the item difficulty level parameter. Other parameters such as the discriminatory power parameter are assumed to be the same for all items and the pseudo-guess parameter is equal to zero (Andayani et al., 2019).

Rasch model analysis can be done using Winsteps software. Analysis that can be done using Winsteps include Person Reability, Item Reliability, Item Fit, Item Measure, and Person Fit. The reliability value of the Rasch Model using Winsteps can be seen by displaying the results from the Output Table main menu, then selecting Table 3.1 Summary Statistics. The reliability value can be seen from the Person Reliability and Item Reliability values that appear. Based on the Winsteps Guide described, Person Reliability can be equated with Classical Test Theory reliability values such as KR-20 and Alpha Cronbach (Boone et al., 2013).

Rasch modeling using Winsteps can be used to see the suitability of items with a model commonly called Item Fit. Item Fit explains whether the item functions normally to take measurements or not. If a question does not fit, it can be indicated that there is a misconception among students about the item. Item Fit on Winsteps can be seen in the output section of Table 2.9 Item Fit Order (Sumintono & Widhiarso, 2015).

Detection of bias on items in the analysis of the Rasch model is shown in the Differential Item Functioning (DIF) function. This is necessary to determine whether the items given have a bias in certain categories of respondents or not. Bias in the items can be identified based on the probability value of the item which is below 5% (Sumintono & Widhiarso, 2014:124).

Currently, there are still not many researchers who use IRT for test analysis in analyzing students' abilities and analyzing the quality of items. Analysis of student abilities and quality of items is still mostly done using the CTT method. Previous research on analyzing the characteristics of items still uses CTT such as the research of Sa'idah et al., (2019) and Sainuddin & Ilyas, (2016).

Research on the quality of items using the Rasch Model has been carried out by Kurniawan (2019), Tabatabaee-Yazdi, (2018), Hasnah (2017), and Dewi Juliah Ratnaningsih & Isfarudi (2013). Likewise with research on the ability of respondents to a test that has been carried out by Moore & Gordon (2014) and Ling et al. (2018).

Some studies emphasis more on comparing the analysis using CTT with IRT as in the study of (Jabrayilov et al., 2016). However, it is still difficult to find research that analyzes the quality of items as well as students' mathematical abilities using the Rasch Model. This study aims to investigate

the quality of the items and students' mathematical abilities using the Rasch Model.

METHODS

This research is a descriptive study that describes the quality of the mathematics questions made by the teacher and the students' abilities in mathematics in one of the State Vocational High Schools in Semarang. The subjects in this study were 126 student response patterns to the test set in the form of multiple choice questions totaling 20 items with five alternative answers in the first semester assessment of mathematics class XII which were collected through documentation techniques.

The questions are made by the supporting teacher by taking into account the existing competency indicators and standards and the quality of the items has not been analyzed. The object of this research is the quality of the items and students' mathematical abilities. The response patterns obtained were analyzed quantitatively using the *Item Response Theory 1 Parameter Logistics approach* or the Rasch Model with the help of *the Winsteps 3.73 program.* Analysis that can be done using *Winsteps* include *Person Reability, Item Reliability, Item Fit, Item Measure, and Person Fit.*

RESULTS AND DISCUSSION

Level of Difficulties (Item Measure)

Item analysis was carried out to find suitable items based on existing criteria. In addition, it was also to determine the level of difficulty of the items tested to students. The order of item difficulty level (item measure) in Table 1 corresponds to the order of items in the Number of Items.

Tabel 1. Level of Difficulties (*Item Measure)*

Number of Item	Measure	Result	Number of Item	Measure	Result
6	1.84	Difficult	3	-0.02	Moderate
13	1.84	Difficult	17	-0.02	Moderate
18	1.06	Difficult	16	-0.02	Moderate
14	0.88	Moderate	5	-0.06	Moderate
8	0.84	Moderate	9	-0.61	Moderate
1	0.41	Moderate	20	-0.71	Moderate
2	0.41	Moderate	7	-1.01	Moderate
12	0.26	Moderate	15	-1.73	Easy
19	0.22	Moderate	10	-1.84	Easy
4	0.14	Moderate	11	-1.95	Easy

The standard deviation in this study is 1.05. The grouping of item difficulty levels can be seen by combining the standard deviation value (1.05) with the logit average which is always 0.00 (Sumintono & Widhiarso, 2015). Values > 0.0 logit + SD are difficult questions, namely items numbered 6, 13, and 18. Values between 0.0 logit + SD to 0.0 logit - SD can be categorized as items with a moderate level of difficulty, so 14 of the 20 items are categorized as a question of moderate difficulty. While the value < 0.0 logit - SD is categorized as an easy question, namely items number 15, 10, and 11.

A high logit value in the Measure column indicates a high level of difficulty. Questions 6 and 13 are the questions with the highest level of difficulty. A total of 37 students out of 126 students answered the question correctly. The question with the lowest level of difficulty is in question number

11. The logit value of the items has the same scale so that it can be used to compare the level of difficulty between one question and another (Sumintono & Widhiarso, 2015). For example, questions number 6 and 12 (+1.84 logit) are compared to item 14 (+0.88 logit), so it can be said that questions number 6 and 12 have twice the difficulty level of question number 14.

Level of Appropriaeness of Items (Item Fit)

Item fit explains whether the items function normally in measuring or not (Sumintono & Widhiarso, 2015). The category of determining fit items is explained by (Boone, 2016). An item is said to be fit if it meets the criteria in Table 2.

Table 2. Question Item Quality Criteria

Standard	Criteria
Outfit Mean Square (MNSQ)	0.5 < MNSQ < 1.5
Outfit Z-Standard (ZSTD)	-2.0 < ZSTD < 2.0
Point Measure Correlation (Pt. Mean Corr)	0.4 < Pt. Measure Corr < 0.85

The result of the item suitability analysis is described in **Table 3.**

Table 3. Level of Appropriaeness of Items (Item Fit)

No. Butir	Outfit		Pt. Mean Corr	No. Butir	Outfit		Pt. Mean Corr
	MNSQ	ZSTD			MNSQ	ZSTD	
7	1.50	1.9	A .22	19	1.02	0.2	j .30
13	1.38	2.7	B .13	11	0.76	-0.5	i .21
10	1.33	0.9	C .17	14	0.96	-0.5	h .38
2	1.14	1.5	D .23	16	0.98	-0.1	g .34
1	1.10	1.1	E .25	8	0.92	-1.1	f .42
3	1.10	0.8	F .26	4	0.87	-1.2	e .42
18	1.07	0.9	G .28	9	0.79	-1.1	d .40
6	1.00	0.1	H .27	20	0.79	-1.0	c .40
15	1.06	0.3	I .12	17	0.87	-1.1	b .43
12	1.03	0.3	J .28	5	0.77	-1.9	a .50

A total of 20 questions have been tested, 19 of which are included in the MNSQ criteria. One item that is not included in the MNSQ criteria is item number 7 with an MNSQ score of 1.5. The ZSTD criteria are met by all items, namely -2.0 < ZSTD < +2.0. Criteria for acceptance of the value of Pt. Measure Corr is 0.4 < Pt. Measure Corr < 0.85 fulfilled by items number 8, 4, 9, 20, 17, and 5. While items that do not include the criteria of Pt. Measure Corr consists of 14 items, namely items number 7, 13, 10, 2, 1, 3, 18, 6, 15, 12, 19, 11, 14, and 16.

The result of the analysis, of the 20 items, only item number 7 is considered unfit so that it needs to be revised or discarded. Compared to other studies, the quality classification of the items analyzed is better than the results of Hasnah's research (2017) where only 9 out of 40 questions or 22.5% items are categorized as good.

Construct Validity

The construct validity of the analyzed items can be seen through Item Dimensionality on Winsteps 3.73. Construct validity can be determined by looking at the

Raw Varience and Unexplned Variance. Raw Variance values can be concluded using the following criteria: Value 20% less good, 20-40% good, 41-60% very good, and > 60% excellent. While the value allowed in the Unexplned Variance is <15% (Widyaningsih & Yusuf, 2018). The results of the construct validity analysis are shown in Figure 1.

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --			Modeled
Total raw variance in observations	=	26.1	100.0%		100.0%
Raw variance explained by measures	=	6.1	23.3%		23.2%
Raw variance explained by persons	=	2.0	7.8%		7.7%
Raw Variance explained by items	=	4.1	15.5%		15.5%
Raw unexplained variance (total)	=	20.0	76.7%	100.0%	76.8%
Unexplned variance in 1st contrast	=	2.7	10.5%	13.7%	
Unexplned variance in 2nd contrast	=	2.2	8.3%	10.9%	
Unexplned variance in 3rd contrast	=	1.9	7.3%	9.5%	
Unexplned variance in 4th contrast	=	1.6	6.2%	8.1%	
Unexplned variance in 5th contrast	=	1.4	5.3%	7.0%	

Figure 1. Construct Validity Analysis Results

The results of the construct validity analysis showed that the Raw variance percentage was 23.3% so it was categorized as good. Raw unexplained variance 11-5 results are 10.5%, 8.3%, 7.3%, 6.2%, and 5.3%, respectively. Overall Raw unexplained variance is below 15%.

Based on the results of the analysis of construct validity using the Rasch Model assisted by Winsteps 3.73, it was found that the instrument analyzed had met the criteria of construct validity with the results of a good category with a percentage of 23.3%. The results obtained by the researcher are not much different from the results of the raw variance of research results from (Widyaningsih & Yusuf, 2018:39) which are also categorized as good with a percentage of 30.8%.

DIF Analysis

A question can be called biased if there is one student with certain characteristics who are more advantaged than the characteristics of other students. In this study, the results were analyzed for gender bias, namely male (L) and female (P). Detection of items infected with DIF can be seen through the output probability (PROB.) in Winsteps 3.73. Items that have a PROB value. below 0.05 indicates that the item is infected with DIF based on gender (Sumintono & Widhiarso, 2014). The results of the DIF analysis are described in Figure 2.

Table 4. DIF Analysis Result

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	.2234	1	.6364	.1102	-.6328	1	i1
2	3.3704	1	.0664	1.7028	.8832	2	i2
2	.1734	1	.6771	.0857	-.7145	3	i3
2	.7937	1	.3730	.3924	-.0969	4	i4
2	2.8087	1	.0938	1.4096	.7286	5	i5
2	1.5664	1	.2107	.7922	.3129	6	i6
2	5.6572	1	.0174	2.9185	1.3816	7	i7
2	.8300	1	.3623	.4136	-.0694	8	i8
2	.2604	1	.6098	.1272	-.5830	9	i9
2	.0035	1	.9526	.0025	-1.3615	10	i10
2	1.4189	1	.2336	.7128	.2450	11	i11
2	.3538	1	.5520	.1745	-.4646	12	i12
2	.6684	1	.4136	.3363	-.1747	13	i13
2	.4858	1	.4858	.2416	-.3287	14	i14
2	2.5597	1	.1096	1.3159	.6747	15	i15
2	1.9510	1	.1625	.9739	.4528	16	i16
2	.1550	1	.6938	.0760	-.7513	17	i17
2	.5760	1	.4479	.2872	-.2503	18	i18
2	.5689	1	.4507	.2819	-.2590	19	i19
2	.1131	1	.7367	.0558	-.8394	20	i20

The results of the DIF analysis in Figure 2. show that one item with a PROB value below 0.05, namely item number 7 with a probability value of 0.0174. The results of the item analysis that are biased towards gender are reinforced by the graph in Figure 2.

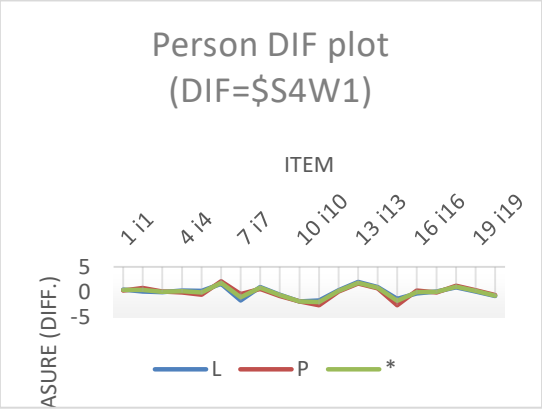

Figure 2. DIF Chart

In Figure 2 above, there are two variations of the classification of respondents, L for male and P for female. There are three lines of different colors. The blue line shows the ability level of the male students and the red line shows the ability level of the female students. While the green line shows the average ability level of all students.

Points that are close to the upper limit such as items number 6, 13, and 18 indicate a high level of problem difficulty. On the other hand, dots at the bottom such as numbers 10, 11, and 15 indicate easy questions. The points of the blue and red lines on the items other than the numbers above coincide with each other so that it can be said that these items do not have any bias in working for both men and women. The item that is considered biased, namely item number 7, indicates that the item is easy for men to do and is considered difficult for women to do. Another item that is considered to be close to bias is number 15, where the item is easy for female testees to do and difficult for male testees to do. However, based on the 0.05 probability criterion, the

item considered to be infected with DIF is item number 7.

Students Ability (Person Measure)

The Person Measure output shows that the average ability of 126 students is 0.85 logit with a standard deviation of 0.79. The mean and standard deviation can be used to classify students' abilities. The average ability of students is used as a reference to classify students' abilities as high, medium, or low. The student's ability criteria are classified based on Table 4.

Table 5. Students' Ability Criteria

Ability	Criteria
High	Measure > Mean + SD
Moderate	Mean - SD < Measure < Mean + SD
Low	Measure < Mean - SD

From the ability of 126 students, there are 27 students with high ability with a Measure value of more than 1.64 logit (0.85+0.79). Then there are 76 students with moderate ability. The remaining 23 students with low ability whose logit value is below 0.06 (0.85-0.79 logit). The ability analysis that has been done shows that 18% of students have low abilities, 60% have moderate abilities, and 22% have high abilities. Medium ability as much as 60% indicates that the majority of students have a fairly good ability to the material being tested.

Person Fit

Rasch model is not only used to classify students' abilities but also used to find out whether there are students with inappropriate response patterns. The pattern of inappropriate responses is the discrepancy of the answers given based on their abilities compared to the ideal model. The level of individual suitability (Person Fit) can be searched with the quality criteria of the items (Item Fit), namely the MNSQ, ZSTD, and Pt criteria. Mean Corr. From 126 students, there were 5 individual responses that were unusual because they fell out of the MNSQ, ZSTD, and Pt criteria. Mean Corr, namely individual responses numbered 121, 88, 1, 103, and 77. The inappropriateness of the response can be seen from the output of the Guttman Scalogram. The output of the Guttman Scalogram has sorted the data based on the students' abilities and the level of difficulty of the questions. The level of difficulty of the questions has been ordered from left to right for the easiest to most difficult questions. The output of the Guttman Scalogram is shown in Figure 3.

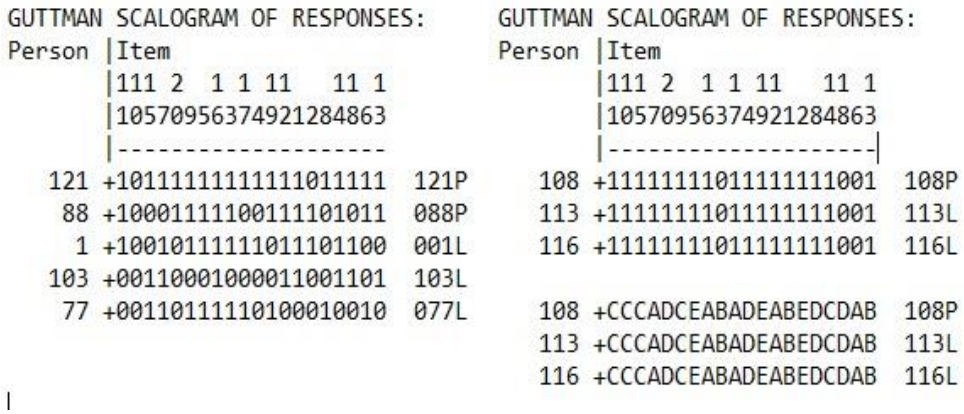

Figure 3. Respone Result of *Guttman Scalogram*

In Guttman Scalogam, it can be seen that respondent number 121 gave a slightly odd response, where he could do the difficult questions but could not do the easier

questions, namely items number 10 and 2. The most unnatural response was response number 77. Respondent number 77 looks random in doing the questions, where the easy questions are not done, but the more difficult questions can be done. The random response pattern from number 77 indicates that response number 77 uses a guessing process (lucky guess) in working on the problem.

The Guttman Scalogram output shows responses that tend to be the same among individuals who work on the problem. Of the 126 students who worked, there were groups that had the same response pattern, so that it could be indicated that there was cheating, namely responses numbered 108, 113, and 116. The indication of cheating was evidenced by the same response on each item numbered question. Moreover, the three respondents are close responses.

Person Reliability and Item Reliability

In addition to analyzing for items and individuals, the analysis of the Rasch model can be used to analyze at the instrument level as a whole. The analysis of the instrument in question includes the average ability of students, comparison of item reliability with individual reliability to the test information function. The standard value of Person Reliability and Item Reliability is > 0.94 special; 0.91 – 0.94 Excellent; 0.81 – 0.90 Good; 0.67 – 0.80 Enough; and < 0.67 Weak (Sumintono & Widhiarso, 2015). The results of the analysis that have been carried out are shown in Figure 4.

SUMMARY OF 126 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	13.0	20.0	.85	.55	.99	.1	1.02	.1
S.D.	2.7	.0	.79	.08	.18	.8	.55	.9
MAX.	19.0	20.0	3.38	1.05	1.64	2.6	4.70	3.0
MIN.	6.0	20.0	-1.03	.50	.64	-2.2	.29	-1.8

REAL RMSE .57 TRUE SD .55 SEPARATION .96 Person RELIABILITY .48
MODEL RMSE .55 TRUE SD .57 SEPARATION 1.03 Person RELIABILITY .51
S.E. OF Person MEAN = .07

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .49

SUMMARY OF 20 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	82.2	126.0	.00	.22	1.00	.0	1.02	.1
S.D.	22.7	.0	1.05	.05	.07	.8	.20	1.2
MAX.	117.0	126.0	1.84	.35	1.13	1.3	1.50	2.7
MIN.	37.0	126.0	-1.95	.19	.86	-1.7	.76	-1.9

REAL RMSE .23 TRUE SD 1.02 SEPARATION 4.38 Item RELIABILITY .95
MODEL RMSE .23 TRUE SD 1.02 SEPARATION 4.44 Item RELIABILITY .95
S.E. OF Item MEAN = .24

UMEAN=.0000 USCALE=1.0000

Figure 4. Instrument Analysis Result

The Mean of Person Measure is 0.85 logit which shows the average value of all students in working on the given problem. This is offset by the Mean of Person Measure

> logit 0.00 indicating the tendency of students' abilities on average to be higher than the level of difficulty of the questions. The good quality of the instrument is supported by the results of the measurement information function which shows the shape of the normal curve. The SD value of 1.05 above 0.00 indicates that respondents have a tendency to be able to answer questions correctly (Nuryanti et al., 2018). Person Reliability value 0.48 and Item Reliability 0.95. The reliability value can be compared with the reliability table.

The Person Reliability value is 0.48 so it is quite weak, while the Item Reliability value is 0.95 so it is very high or special. The value of person reliability is lower than item reliability, indicating that the consistency of student answers is weak, but the quality of the items in the instrument is very high. A good instrument can also be proven by the test information function graph in Figure 5.

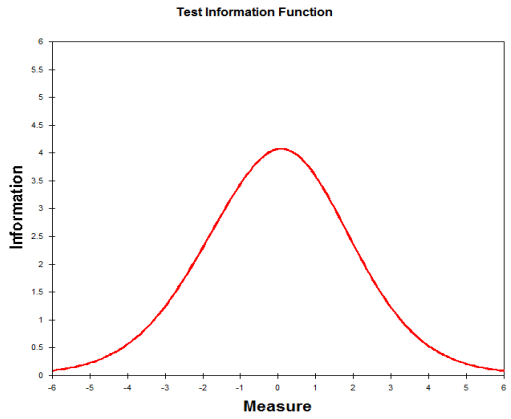

Figure 5. Test Information Function

The test information function in Figure 5 shows a normal curve. The X-axis shows the level of student's ability to work on the questions, while the Y-axis explains the magnitude of the information function. At a low level of ability, namely on the left side of the curve, it shows that the information obtained is quite low, as well as a high level of ability. The information obtained by the measurement is very high at the average ability level of 0.00 logit. Thus, the graph shows that the items are suitable for knowing the level of students' abilities.

CONCLUSION

From the research that has been carried out, it was found that of the 20 items tested, 19 of them functioned normally in measuring students' abilities, while one item was considered inappropriate for measuring students' abilities as well as gender bias. A total of 3 questions are categorized as difficult questions, 14 items are in the medium category, and 3 questions are in the easy category. As many as 126 students who gave responses, there were 27 students with high abilities, 76 students with moderate abilities, and the remaining 23 students with low abilities. Overall, the instrument can be said to be good with item reliability of 0.95 and a curve graph that forms a normal curve.

The ability of students who are quite weak with a person reliability of 0.48 becomes a note for teachers to strengthen students' abilities, especially the basics of mathematics so that students' abilities can get better.

REFERENCES

Andayani, A., Purwanto, & Ramalis, T. R. (2019). Kajian implementasi teori respon butir dalam menganalisis instrumen tes materi fisika. *Prosiding Seminar Nasional Fisika 5.0*, *1*(1), 37–42.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4), 1–7.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch Analysis in the Human Sciences*. Springer Science & Business Media.

Dewi Juliah Ratnaningsih, & Isfarudi. (2013). Analisis Butir Tes Objektif Ujian Akhir Semester Mahasiswa Universitas Terbuka Berdasarkan Teori Tes Modern. *Jurnal*

Pendidikan Terbuka Dan Jarak Jauh, *14*(2), 98–109.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri dan Paralel Melalui Teori Tes Klasik dan Model Rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11–19.

Fauziana, A., Wulansari, A. D., I, S. D. N. K., & Ponorogo, I. (2021). *Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch*. *6*, 10–19.

Fernanda, J. W., & Hidayah, N. (2020). Analisis Kualitas Soal Ujian Statistika Menggunakan Classical Test Theory dan Rasch Model. *Square : Journal of Mathematics and Mathematics Education*, *2*(1), 49.

Hasnah. (2017). Analisis Kualitas Soal Matematika Ujian Sekolah Kelas XII IPA SMA Negeri di Watansoppeng Berdasarkan Teori Respon Butir. *PEP Educational Assessment*, *1*(1), 27–33.

Hazlita, S., Zulkardi, & Darmawijoyo. (2014). Pengembangan Soal Penalaran Model TIMSS Konteks Sumatera Selatan di Kelas IX SMP. *Jurnal Kreano*, *5*(November), 170–179.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*(8), 559–572.

Kurniawan, D. D. (2019). Analisis Butir Soal Ujian Akhir Semester Matematika dengan Teori Respon Butir. *Briliant: Jurnal Riset Dan Konseptual*, *4*(20), 215–224.

Ling, M., Pang, V., & Ompok, C. C. (2018). Measuring Change in Early Mathematics Ability of Children Who Learn Using Games: Stacked Analysis in Rasch Measurement. *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings*.

Moore, M., & Gordon, P. C. (2014). Reading Ability and Print Exposure: Item Response Theory Analysis of The Author Recognition Test. *Behavior Research Methods*, *47*(4), 1095–1109.

Nuryanti, S., Masykuri, M., & Susilowati, E. (2018). Analisis Iteman dan Model Rasch pada Pengembangan Instrumen Kemampuan Berpikir Kritis Peserta Didik Sekolah Menengah Kejuruan. *Jurnal Inovasi Pendidikan IPA*, *4*(2), 224–233.

Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy : Jurnal Pendidikan Islam*, *7*(1), 61–70.

Sa'idah, N., Yulistianti, H. D., & Megawati, E. (2019). Analisis Instrumen Tes Higher Order Thinking. *Jurnal Pendidikan Matematika*, *13*(1), 41–54.

Sainuddin, S., & Ilyas, M. (2016). Kakteristik Butir Tes Matematika pada Tes Buatan MGMP Matematika Kota Palopo Berdasarkan Teori Klasik. *Pedagogy: Jurnal Pendidikan Matematika*, *1*(1), 125–139.

Sumintono, B. (2018). Rasch Model Measurements as Tools in Assesment for Learning. *1st International Conference on Education Innovation*, *173*(Icei 2017), 38–42.

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*. Penerbit Trim Komunikata.

Tabatabaee-Yazdi, M. (2018). Development and Validation of a Teacher Success Questionnaire Using the Rasch Model. *International Journal of Instruction*, *11*(2), 129–144.

Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education*, *58*(8), 922–933.

Widyaningsih, S. W., & Yusuf, I. (2018). Analisis Soal Modul Laboratorium Fisika Sekolah I Menggunakan Racsh Model. *Gravity: Jurnal Ilmiah Penelitian Dan Pembelajaran Fisika*, *4*(1), 33–46.