BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0151–0162
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0151-0162

ITEM ANALYSIS OF HIGH SCHOOL SPECIALIZATION MATHEMATICS EXAM QUESTIONS WITH ITEM RESPONSE THEORY APPROACH

Lovieanta Arriza [1*], Heri Retnawati[2], Rizki Tika Ayuni[3]

*[1,2,3]Educational Research and Evaluation Program, Graduate School, Universitas Negeri Yogyakarta
Karangmalang Campus, Colombo Street No.1, Yogyakarta, 55281, Indonesia*

*Corresponding author's e-mail: * **lovieanta0341pasca.2022@student.uny.ac.id***

ABSTRACT

Article History:

Received: 1st August 2023
Revised: 20th November 2023
Accepted: 19th December 2023

Keywords:

Item Analysis;
Item Response Theory;
Rasch Model;
Dichotomus.

Analysis of item characteristics on test instruments is carried out to determine high-quality items. This study aims to describe the parameters of specialized high school mathematics test items using the IRT approach. It is an exploratory, descriptive study employing a quantitative approach. The research subjects were 36 students of grade XI high school who took the specialization mathematics subject. Response data with dichotomous scoring were analyzed using the IRT approach with the R program to obtain information about item parameters and student ability. The results of the model fit test showed that most of the specialization mathematics exam items fit the Rasch model. The results showed that all items met the criteria of good quality because they had good difficulty parameters. Relatively, the test items were suitable for students with abilities between -2.6 and 2.8 logits. This estimation is also supported by the TIF with a maximum value of 3.049 at 0.08 logit ability and SEM of 0.541. Test items that have been proven to be of high quality can be used as examples in both teaching and diagnostic assessments. Further research could consider the discrimination parameter when analyzing the characteristics of the questions.

How to cite this article:

L. Arriza, H. Retnawati and R. T. Ayuni., "ITEM ANALYSIS OF HIGH SCHOOL SPECIALIZATION MATHEMATICS EXAM QUESTIONS WITH ITEMS RESPONSE THEORYAPPROACH", *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0151-0162, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

One of the main and integral components of the learning process is assessment [1], [2]. Improving the quality of education can be achieved through improving the quality of learning and the quality of the assessment system. The assessment process will provide information on the learning outcomes that have been carried out [3], [4]. In addition, assessment also provides feedback on student learning progress [5], [6]. In educational assessment, tools are needed in the form of assessment instruments [7]. There are two types of assessment instruments, namely tests to measure student learning achievement, intelligence, aptitude, skills, and nontests. Most instruments contain one or a set of items that aim to measure the knowledge and skill domain or task domain that represents the student's ability [8].

The form of test instrument commonly used in educational assessment and measurement is multiple choice. A multiple-choice test is a set of tests that contain questions with two or more alternative answers, but only one answer is correct, and the test takers have to choose one correct answer [9]. Multiple-choice tests are widely used because of their ease and effectiveness in measuring students' knowledge or skill domains [10], [11], [12]. Multiple-choice questions also allow teachers to quickly measure a wide range of knowledge, skills, and competencies across disciplines and fields including student's ability to understand concepts and principles, make judgments, draw conclusions, give reasons, complete statements, interpret data, and apply information [13], [14].

The test instruments that have been used still need to be tested empirically for quality. Data from the test results will be analyzed to obtain evidence about the characteristics of the items concerned. Then from the results of the analysis of the characteristics of these items will be obtained the basis for making the necessary revisions. The quality of the instrument requires validity, reliability, and good item parameters [15]. Conclusions about the characteristics of the items on a test will lead to decisions about whether or not the item should be used, whether it should be discarded, whether it can still be revised, or whether it has indeed met the requirements of a good item. The systematic work procedure for evaluating all test items based on empirical data is called item analysis [16]. Item analysis is a key step in evaluating tests in the field of educational measurement [17].

In educational measurement, there are two types of approaches that can be used to analyze instrument quality, namely classical test theory and modern theory [18]. Although it is called classical test theory, it is still used today. The purpose of both test theories is the same, which is to find the most appropriate way to obtain a pure score that reflects the actual ability of the participants. However, classical test theory is considered to have weaknesses. The main weakness of classical test theory is that the characteristics of items and examinees cannot be separated. Item parameters are highly dependent on the ability of participants, and vice versa [19], [20]. In other words, the analysis approach with classical test theory cannot be used because the assessment results are highly dependent on the group of test takers.

Item response theory (IRT) has emerged as a solution to current analytical approaches in educational measurement [21], [22]. Item response theory is a modern psychometric framework that provides several beneficial traits compared to classical test theory [23]. Using IRT, the difficulty, discrimination and efficiency of the item information function can be evaluated [24]. In IRT, the probability of a subject answering an item correctly depends on the subjects' ability and item characteristics [25]. This means that test takers with high ability will have a greater probability of answering correctly when compared to participants who have low ability [22], [26].

The application of IRT in analyzing the quality of instrument items has been widely published. As revelaed in Malaspina and Arias [27], which describes the characteristics of items in an early mathematics test instrument using the Item Response Theory (IRT) approach. The results indicate that the item characteristics align with the student sample and do not contain bias. Another study conducted by Ramadhani et al.[28] analyzed items related to statistical reasoning in the context of ethnomathematics for junior high school students. This finding further strengthens the notion that IRT analysis provides consistent information about item characteristics that remains unchanged regardless of students' varying abilities. In addition to these two studies, there is also another research examining the characteristics of mathematics items, specifically focusing on the topic of rational numbers. Khairani and Shamsuddin [29] conducted this study, concluding that the formulated items have good quality and serve as a guide for teachers to use high-quality items in classroom assessments.

The specialization mathematics subject taught in high schools today is sufficiently intriguing to be chosen as the object of analysis. Based on previous research, no research has examined the quality of specialization mathematics exam questions used in schools. Specialization mathematics is a subject taken by students with a particular interest in mathematics, especially those who intend to continue their studies in mathematics-related fields, thereby increasing their chances of working in mathematics-related fields. This underscores the importance of analyzing the characteristics of questions that measure students' abilities in specialization mathematics in secondary schools.

Based on the discussions with the high school mathematics teacher, the information obtained is that so far, no analysis of test items has been carried out by the teacher, either by applying classical test theory or IRT. The test questions developed by teachers are directly used in the assessment process without determining the characteristics of each item. This also applies to specialization mathematics exam questions. In fact, it is important for teachers to know the characteristics of test items so that they are in accordance with the measurement objectives [30].

Based on preliminary studies and relevant research publications that have been stated previously, the application of IRT in analyzing item parameters will be more profitable than classical test theory, both in determining the characteristics of test items and estimating student abilities. Thus, this research aims to describe the characteristics of specialization mathematics exam items using the IRT approach developed by teachers. This study hopes to determine the quality of mathematics specialization exam questions used in schools. In addition to being used for final assessment, high-quality exam questions can be a reference in preparing other questions and are beneficial for teachers in providing examples of problems in classroom learning. This is very important considering that specialization mathematics is a subject attended by students who have a particular interest in exploring mathematics.

2. RESEARCH METHODS

This research is exploratory, descriptive research with a quantitative approach that aims to describe the characteristics of specialization mathematics exam questions developed by teachers. The research subjects were 36 class XI high school students consisting of 10 male students and 26 female students. The sample was chosen because they were students taking specialization mathematics subjects. The test instrument used is specialization mathematics questions consisting of 15 multiple choice (dichotomy) questions with five response categories. Based on the test instrument, student answer results are obtained and collected through the documentation method. The instrument validation aspects analyzed include: a) prerequisite tests for the IRT model, namely unidimensionality and local independence tests, b) model fit testing, c) estimation of item difficulty levels, and d) measurement information function.

Before the data is analyzed with the IRT approach, there are two assumptions that should be fulfilled. Ahmad and Mokshein stated that the assumptions of unidimensionality and local independence should be tested before conducting IRT-based item analysis [31]. Conducting a unidimensionality test in IRT is very important to prove that the question instrument only measures one dimension of ability [32], [33]. The unidimensionality assumption test can be proven through factor analysis by reviewing the eigenvalues in the inter-item covariance variance matrix or can be seen in the eigenvalue scree plot, which shows one dominant component [34], [35]. The unidimensionality test in this case is proven through factor analysis using the R program. Factor analysis is carried out by first conducting a feasibility test analysis, namely the KMO-MSA test (Kaiser-Meyer-Olkin measure of sampling adequacy) and Bartlett's test. The KMO-MSA test aims to see the adequacy of the sample, while Bartlett's Test serves to prove the homogeneity of the data. If the (KMO)-MSA value is > 0.5 and Bartlett's significant test is < 0.05, the unidimensionality test can continue [36].

The unidimensional assumption is proven through principal component analysis (PCA) by counting eigenvalue (λm) of covarince matrix using Equation (1) as follows [37]:

$$C_x v_m = \lambda_m v_m \tag{1}$$

The next assumption test is local independence. According to Hambleton et al [19], local independence is mathematically expressed by the Equation (2):

$$P(u_1, u_2, \dots, u_n|\theta) = P(u_1|\theta), P(u_2|\theta) \dots P(u_n|\theta)$$

$$= \prod_{i=1}^{n} P(u_i|\theta) \quad (2)$$

Description:

i : 1, 2, 3, ... n

n : Number of test item

$P(u_i|\theta)$: The probability of test takers who have the ability θ can answer item number-correctly

$P(u_1, u_2, \ldots, u_n|\theta)$: The probability of test takers who have the ability θ can answer item number-*i* to number-*n* correctly

After the IRT assumptions are met, the next analysis is to test model suitability. *Chi-square* (χ^2) method is one of statistical test of model suitability which used to test latent [38]. Equation (3) below is *Chi-square* (χ^2).

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

Description:

X^2 : *Chi-square* distribution

O_i : Observation value number-*i*

E_i : Expectation value number-*i*

Item parameter estimation is carried out after determining a suitable model. The formula used in this analysis refers to the Rasch model which contains item difficulty level parameters. Below is Equation (4) for estimating item parameters with the Rasch model [39].

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (4)$$

Description:

$P_i(\theta)$: The probability of test takers who have the ability θ are randomly selected can answer item order-*i* correctly

θ : Ability level

b_i : Difficulty level item number-*i*

e : Natural numbers that are close in value 2,718

The results of estimating item parameters using the Rasch model above are then strengthened by the Item Information Function (IIF). Mathematically, the item information function satisfies Equation (5) as follows [26]:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta) Q_i(\theta)} \quad (5)$$

Description:

i : 1, 2, 3, ..., n

$I_i(\theta)$: Information function number-*i*

$P_i(\theta)$: The probability that a participant with ability θ will answer item number-*i* correctly

$P_i'(\theta)$: Derivative function $P_i(\theta)$ to θ

$Q_i(\theta)$:The probability that a participant with ability θ will answer item number-*i* incorrectly

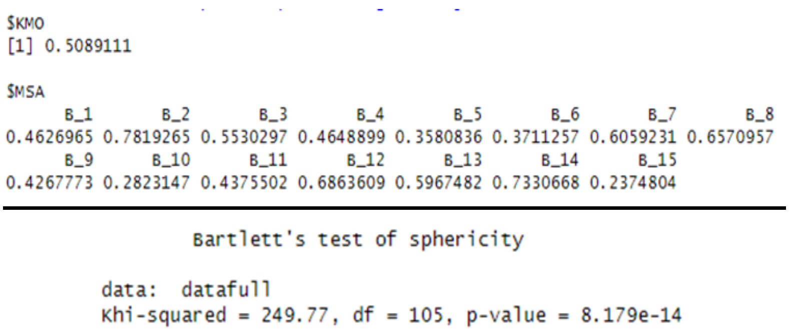
The criteria for a valid instrument based on several aspects can be seen in Table 1.

Table 1. Valid Test Criteria are Seen from Various Aspects and Criteria [40]

The validity aspect of the item	Criteria
Unidimensional test	There is only one dominant factor in the specialization mathematics exam questions through scree plot factor analysis
Local independence	Can be known by proving the unidimensional assumption
Model fit testing	$\chi2 > \chi2_{\text{tabel}}$ $sig > \alpha$ (0,05)
Index difficulty item	Good: $-2 \leq b_i \leq 2$ Not good: $b_i > 2$; $b_i < -2$
Person ability (wright map)	All levels of difficulty of the questions are within the student's ability domain
Test Information Function	The information function test has the maximum value in the student's ability domain

3. RESULTS AND DISCUSSION

Based on the preliminary analysis conducted, the KMO-SMA and Barlett's values were obtained as a measure of sample adequacy, with the output presented in Figure 1 below.

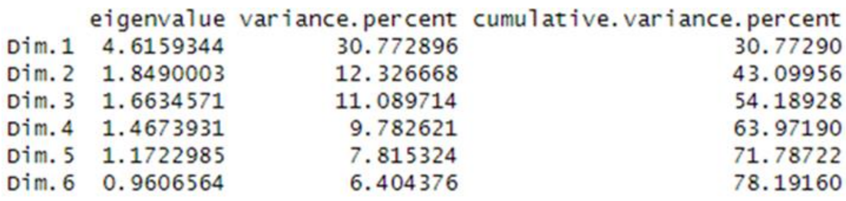


```
$KMO
[1] 0.5089111

$MSA
      B_1       B_2       B_3       B_4       B_5       B_6       B_7       B_8
0.4626965 0.7819265 0.5530297 0.4648899 0.3580836 0.3711257 0.6059231 0.6570957
      B_9      B_10      B_11      B_12      B_13      B_14      B_15
0.4267773 0.2823147 0.4375502 0.6863609 0.5967482 0.7330668 0.2374804
```

```
              Bartlett's test of sphericity

       data:  datafull
       Khi-squared = 249.77, df = 105, p-value = 8.179e-14
```

Figure 1. KMO and Bartlett's test

Based on Figure 1, it could be seen that the KMO-MSA value was 0.508 and the significant Bartlett's test was 0.000. This indicates that the 36 samples used for IRT analysis have met the requirements of sample adequacy and the data are homogeneous so that factor analysis can be carried out. The unidimensionality assumption test can be proven through factor analysis by reviewing the eigenvalue in the inter-item covariance variance matrix or can be seen in the eigenvalue scree plot which shows one dominant component. The results of data processing for principal component analysis with the R program can be seen in the eigenvalue section in Figure 2 below.



```
      eigenvalue variance.percent cumulative.variance.percent
Dim.1  4.6159344        30.772896                    30.77290
Dim.2  1.8490003        12.326668                    43.09956
Dim.3  1.6634571        11.089714                    54.18928
Dim.4  1.4673931         9.782621                    63.97190
Dim.5  1.1722985         7.815324                    71.78722
Dim.6  0.9606564         6.404376                    78.19160
```

Figure 2. Test Results of The Main Components

Based on Figure 2, it can be found that the response data contains 5 components. From the five components, component 1 was the most dominant component because it has the largest eigenvalue of 4.615. The difference between the eigenvalues of components 1 and 2 is 2.7669341. Meanwhile, the difference between components 2 and 3 is 0.1855432. The eigenvalue differences for the subsequent components remain less than the differences between the eigenvalues of components 1 and 2. This further indicates a significant decrease in the eigenvalue differences, especially between component 1 and the other components. This factor value can then be presented in the scree plot in Figure 3.

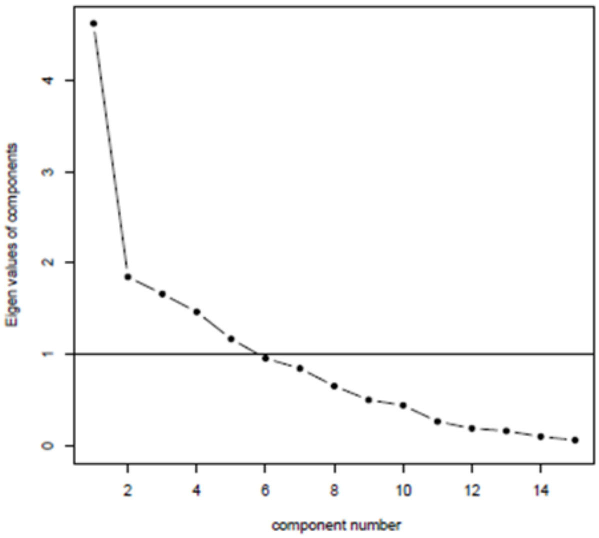


Figure 3. Scree Plot of Factor Analysis

Figure 3 above is the scree plot result of the factor analysis. The dots in the figure show the number of dimensional factors in the math specialization exam questions. The connecting lines between the points show the distance and slope of the differences between factors. The scree plot of the factor analysis shows a very sharp drop between factor 1 and factor 2, and the eigenvalues then start to skew at factor 3 so that the scree plot almost forms a right angle. This indicates that there was only one dominant factor in the math specialization exam questions. The results of this analysis are in accordance with the assumptions in the IRT approach where a set of questions or tests only measure one latent trait so that the assumption of unidimensionality is met [19].

The next assumption test is local independence. In its proof, this local independence can be fulfilled if the participant's answer to an item does not affect the participant's answer to another item [41], [22]. Local independence can be known by proving the unidimensionality assumption [19]. In this analysis, it can be seen in Figure 3 that the assumption of unidimensionality has been met so that the local independence test has also been met.

Furthermore, the model fit was assessed using various indicators. One of them was to compare the output of several different model fit indices, Akaike information criterion (AIC), Bayesian information criterion (BIC), loglikelihood which can be seen in Table 2.

Table 2. Comparison of Model Fit Tests

Model	AIC	BIC	loglikelihood
Rasch	666.99	692.33.00	-317.50
2PLM	675.71	723.21.00	-307.85
3PLM	698.95	770.21.00	-304.48

Table 2 provides information on the values obtained from the calculation of AIC, BIC, and loglikelihood for each model. In order to determine which model best fits the test data, the selection criteria are seen from the smallest value in each model. The smaller the fit index value, the better the model fits the data [42]. Table 3 shows that the Rasch model provides the smallest value among other models in each AIC, BIC, and loglikelihood.

Model fit testing can also be done using statistical methods by determining the chi-square of each item on each IRT logistic parameters. This was done by calculating the chi-square ($\chi 2$) value and then comparing it with the chi-square ($\chi 2$) value from the table, or by reviewing the probability (significance) value. Items are said to fit the model if the calculated chi-square value was smaller than the chi-square ($\chi 2$) table or sig value > α (0.05) [43]. The results of the fit test for each item in the Rasch, 2 PL and 3 PL models were presented in Table 3.

Table 3. The fit of each item in the Rasch, 2PL and 3PL models

Item	Rasch Model		2PL		3PL	
	Sig.	Category	Sig.	Category	Sig.	Category
B_1	0.149	Fit	0.187	Fit	0.700	Fit
B_2	0.717	Fit	0.375	Fit	0.219	Fit
B_3	0.369	Fit	0.379	Fit	0.018	Fit
B_4	0.497	Fit	0.490	Fit	0.215	Fit
B_5	0.513	Fit	0.880	Fit	0.467	Fit
B_6	0.512	Fit	0.437	Fit	0.677	Fit
B_7	0.418	Fit	0.645	Fit	0.697	Fit
B_8	0.262	Fit	0.390	Fit	0.073	Fit
B_9	0.417	Fit	0.494	Fit	0.366	Fit
B_10	0.282	Fit	0.462	Fit	0.427	Fit
B_11	0.056	Fit	0.046	Not Fit	0.079	Not Fit
B_12	0.116	Fit	NaN	NaN	NaN	NaN
B_13	0.674	Fit	0.706	Fit	0.696	Fit
B_14	0.212	Fit	0.120	Fit	0.037	Fit
B_15	0.188	Fit	0.208	Fit	0.126	Fit

Based on Table 3, it can be seen that all items were suitable for the Rasch model, 13 items were suitable for the 2PL model and the suitability for the 3PL model was also 13 items. Based on the percentage, the suitability with the Rasch model is the greatest compared to 2PL and 3 PL. Therefore, it can be concluded based on this analysis that the analysis of the instrument of specialization mathematics exam questions was suitable for the Rasch model.

Rasch is a measurement model developed by mathematician George Rasch from Denmark. Analysis with the Rasch model is determined based on the level of item difficulty and participant ability simultaneously. The probability of answering an item correctly is differentiated based on item difficulty and individual ability. People with low ability should not be able to correctly answer items that have a high level of difficulty [44]. In this case, the scoring used to be analyzed with the Rasch model was dichotomous data on the results of the specialization mathematics exam, namely the wrong answer was given a score of 0 and the correct answer was given a score of 1.

The next analysis was the estimation of item parameter values with reference to the Rasch model, namely the item difficulty parameter. In the analysis of item characteristics using the Rasch model, an item is said to have good quality if it has a difficulty index (b) ranging from $-2 \leq b \leq 2$ [19]. The overall results of item parameter estimation using the Rasch model were presented in Table 4.

Table 4. Results of Item Parameter Estimation with The Rasch Model

Item	Index Difficulty	Criteria
1	0.159	Good
2	-0.137	Good
3	0.308	Good
4	0.611	Good
5	0.159	Good
6	0.308	Good
7	-0.438	Good
8	-0.137	Good
9	-1.253	Good
10	1.445	Good
11	-0.591	Good
12	-0.438	Good

Item	Index Difficulty	Criteria
13	-0.287	Good
14	0.611	Good
15	1.091	Good

Based on Table 4, it was obtained that all items of the specialization mathematics exam with dichotomous scoring had good item difficulty parameters. This information proved that 100% of all test items were able to describe the function of students' abilities. Where students who have high abilities will find it easy to work on the items, conversely students who have low abilities will find it difficult to answer the items.

The next analysis showed information that the ability of students who answered the specialization math exam items was in theta (θ) -1.955 to 2.093 logit with an average theta of 0.00. Based on this data, it could be concluded that it was natural that many items were classified as moderate for students because their abilities were relatively moderate or average abilities. The estimation of participants' abilities can also be seen in Figure 4.

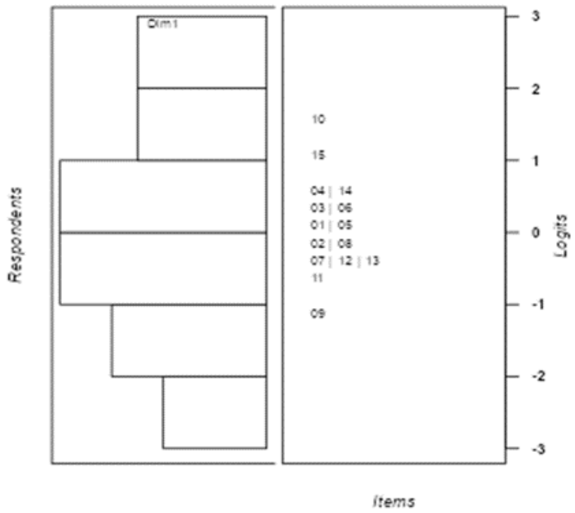


Figure 4. Student ability estimation graph

The subsequent analysis of the information function yields values that are inversely related to the Standard Error of Measurement (SEM), as presented in Figure 5.

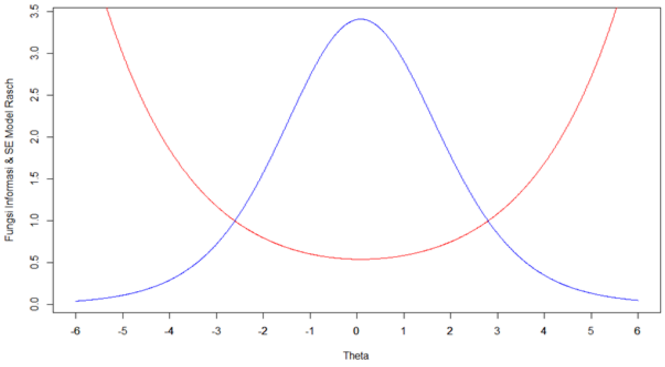


Figure 5. Information Function Graph of Test and SEM

Figure 5 above is the result of the Test Information Function (TIF) and SEM analysis. The blue colored curve line is the TIF graph that has a high point of 3.049 at theta (θ) 0.08 logit. Meanwhile, the red line which is inversely proportional to the TIF is the SEM graph with the lowest point of 0.541. This meant that the specialization mathematics exam items produced optimal information when used by students with 0.08 logit ability. The TIF and SEM curves intersected at theta -2.6 and 2.8 which means that the test was overall suitable for students with abilities between -2.6 and 2.8 logits. This range also indicated that the specialization math exam questions were able to measure the ability of students with a fairly wide range [45]. This is an important note considering that the abilities of students in all schools are not at the same level. Therefore, test questions prepared by teachers must be able to facilitate various levels of student ability **[46]**, **[47]**.

4. CONCLUSIONS

Based on the results of the study, it can be concluded that the test sample data has met the IRT assumptions. The results of the model fit test analysis show that all items of the specialization mathematics exam are suitable for estimation with the Rasch model compared to the 2PL and 3PL models. From the results of estimating item parameters with the Rasch model, all items meet the criteria for a good difficulty index because the dominant ability of the participants measured is classified as moderate. The ability of students who answered the math specialization exam items was in the theta (θ) -1.955 to 2.093 logit with an average theta of 0.00. The results of estimating item parameters and student ability are supported by the Test Information Function (TIF) with a maximum value of 3.049 at 0.08 logit ability. This is reinforced by the SEM curve, which is inversely proportional to the TIF with the lowest point of 0.541. This provides information that the specialization math exam items produce optimal information when used by students with a logit ability of 0.08.

REFERENCES

[1] S. Ndiung and M. Jediut, “Pengembangan instrumen tes hasil belajar matematika peserta didik sekolah dasar berorientasi pada berpikir tingkat tinggi,” *Premiere Educandum : Jurnal Pendidikan Dasar dan Pembelajaran*, vol. 10, no. 1, p. 94, Jun. 2020, doi: 10.25273/pe.v10i1.6274.

[2] L. W. T. Schuwirth and C. P. M. Van der Vleuten, “Programmatic assessment: From assessment of learning to assessment for learning,” *Med Teach*, vol. 33, no. 6, pp. 478–485, Jun. 2011, doi: 10.3109/0142159X.2011.565828.

[3] A. N. Castleberry, E. F. Schneider, M. H. Carle, and C. D. Stowe, “Development of a summative examination with subject matter expert validation,” *Am J Pharm Educ*, vol. 80, no. 2, pp. 1–9, Mar. 2016, doi: 10.5688/ajpe80229.

[4] S. Heeneman, A. Oudkerk Pool, L. W. T. Schuwirth, C. P. M. van der Vleuten, and E. W. Driessen, “The impact of programmatic assessment on student learning: Theory versus practice,” *Med Educ*, vol. 49, no. 5, pp. 487–498, May 2015, doi: 10.1111/medu.12645.

[5] L. Ivanjek *et al.*, “Development of a two-tier instrument on simple electric circuits,” *Phys Rev Phys Educ Res*, vol. 17, no. 2, p. 020123, Sep. 2021, doi: 10.1103/PhysRevPhysEducRes.17.020123.

[6] G. Svensäter and M. Rohlin, “Assessment model blending formative and summative assessments using the SOLO taxonomy,” *European Journal of Dental Education*, vol. 27, no. 1, pp. 149–157, Feb. 2023, doi: 10.1111/eje.12787.

[7] D. Desilva, I. Sakti, and R. Medriati, “Pengembangan instrumen penilaian hasil belajar fisika berorientasi hots (higher order thinking skills) pada materi elastisitas dan hukum hooke,” *Jurnal Kumparan Fisika*, vol. 3, no. 1, pp. 41–50, Apr. 2020, doi: 10.33369/jkf.3.1.41-50.

[8] T. M. Haladyna and M. C. Rodriguez, *Developing and validating test items*. Routledge, 2013. doi: 10.4324/9780203850381.

[9] A. M. Andrés and J. D. L. Castillo, “Multiple choice tests: Power, length and optimal number of choices per item,” *British Journal of Mathematical and Statistical Psychology*, vol. 43, no. 1, pp. 57–71, May 1990, doi: 10.1111/j.2044-8317.1990.tb00926.x.

[10] O. N. Bakytbekovich *et al.*, “Distractor analysis in multiple-choice items using the rasch model,” *International Journal of Language Testing*, vol. 13, pp. 69–78, 2023, doi: 10.22034/IJLT.2023.387942.1236.

[11] D. Briggs, A. Alonzo, C. Schwab, and M. Wilson, “Diagnostic assessment with ordered multiple-choice items,” *Educational Assessment*, vol. 11, no. 1, pp. 33–63, Feb. 2006, doi: 10.1207/s15326977ea1101_2.

[12] S. Lions, C. Monsalve, P. Dartnell, M. P. Blanco, G. Ortega, and J. Lemarié, “Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review,” *Applied Measurement in Education*, vol. 35, no. 2, pp. 133–152, Apr. 2022, doi: 10.1080/08957347.2022.2067539.

[13] M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang, “Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review,” *Rev Educ Res*, vol. 87, no. 6, pp. 1082–1116, Dec. 2017, doi: 10.3102/0034654317726529.

[14] M. M. Wooten, A. M. Cool, E. E. Prather, and K. D. Tanner, “Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory,” *Physical Review Special Topics - Physics Education Research*, vol. 10, no. 2, p. 020103, Jul. 2014, doi: 10.1103/PhysRevSTPER.10.020103.

[15] D. Saepuzaman, H. Haryanto, , Edi Istiyono, H. Retnawati, and Y. Yustiandi, “Analysis of items parameters on work and energy subtest using item response theory,” *Jurnal Pendidikan MIPA*, vol. 22, no. 1, pp. 1–9, 2021, doi: 10.23960/jpmipa/v22i1.pp1-9.

[16] R. Pratiwi, S. Reflianti, S. Antini, and A. Walid, “Analysis of item difficulty index for midterm examinations in junior high schools 5 Bengkulu City,” *Asian Journal of Science Education*, vol. 3, no. 1, pp. 12–18, Apr. 2021, doi: 10.24815/ajse.v3i1.18895.

[17] H. Guo, R. Lu, M. S. Johnson, and D. F. McCaffrey, “Alternative methods for item parameter estimation: From CTT to IRT,” *ETS Research Report Series*, vol. 2022, no. 1, pp. 1–16, Dec. 2022, doi: 10.1002/ets2.12355.

[18] R. Bahar, E. Istiyono, W. Widihastuti, S. Munadi, Z. Nuryana, and S. Fajaruddin, “Analisis karakteristik soal ujian sekolah hasil musyawarah guru matematika di Tasikmalaya,” *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, vol. 10, no. 4, pp. 2660–2674, Dec. 2021, doi: 10.24127/ajpm.v10i4.4359.

[19] R. K. Hambleton, Hariharan. Swaminathan, and H. Jane. Rogers, *Fundamentals of item response theory*. Sage Publications, 1991.

[20] D. Saepuzaman, E. Istiyono, H. Haryanto, H. Retnawati, and Y. Yustiandi, “Analisis karakteristik butir soal fisika dengan pendekatan IRT penskoran dikotomus dan politomus,” *Radiasi : Jurnal Berkala Pendidikan Fisika*, vol. 14, no. 2, pp. 62–75, Sep. 2021, doi: 10.37729/radiasi.v14i2.1200.

[21] M. A. Khalaf and E. M. N. Omara, “Rasch analysis and differential item functioning of English language anxiety scale (ELAS) across sex in Egyptian context,” *BMC Psychol*, vol. 10, no. 1, p. 242, Oct. 2022, doi: 10.1186/s40359-022-00955-w.

[22] Y. Liu and A. Maydeu-Olivares, “Local dependence diagnostics in IRT modeling of binary data,” *Educ Psychol Meas*, vol. 73, no. 2, pp. 254–274, Apr. 2013, doi: 10.1177/0013164412453841.

[23] G. Fergadiotis *et al.*, “Item response theory modeling of the verb naming test,” *Journal of Speech, Language, and Hearing Research*, vol. 66, no. 5, pp. 1718–1739, May 2023, doi: 10.1044/2023_JSLHR-22-00458.

[24] M. Brucato, A. Frick, S. Pichelmann, A. Nazareth, and N. S. Newcombe, “Measuring spatial perspective taking: Analysis of four measures using item response theory,” *Top Cogn Sci*, vol. 15, no. 1, pp. 46–74, Jan. 2022, doi: 10.1111/tops.12597.

[25] J. Kean, E. F. Bisson, D. S. Brodke, J. Biber, and P. H. Gross, “An introduction to item response theory and rasch analysis: Application using the eating assessment tool (EAT-10),” *Brain Impairment*, vol. 19, no. 1, pp. 91–102, Mar. 2018, doi: 10.1017/BrImp.2017.31.

[26] H. Retnawati, *Teori respons butir dan penerapannya*. Nuha Medika, 2014.

[27] M. Malaspina and B. Arias, “A Rasch modeling approach for measuring young children’s informal mathematics in Peru,” *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 18, no. 9, pp. 1–13, Aug. 2022, doi: 10.29333/ejmste/12303.

[28] R. Ramadhani, S. Saragih, and E. E. Napitupulu, “Exploration of students’ statistical reasoning ability in the context of ethnomathematics: A study of the Rasch model,” *Mathematics Teaching Research Journal*, vol. 14, no. 1, pp. 138–168, 2022.

[29] A. Z. Khairani and H. Shamsuddin, “Application of rasch measurement model in developing calibrated item pool for the topic of rational numbers,” *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 17, no. 12, pp. 1–11, Dec. 2021, doi: 10.29333/ejmste/11426.

[30] T. T. Semiun and F. D. Luruk, “The quality of an english summative test of a public junior high school, Kupang-NTT,” *English Language Teaching Educational Journal*, vol. 3, no. 2, p. 133, Sep. 2020, doi: 10.12928/eltej.v3i2.2311.

[31] H. Ahmad and S. E. Mokshein, “Is 3pl item response theory an appropriate model for dichotomous item analysis of the anatomy & physiology final examination?,” *Malaysian Science & Mathematics Education Journal*, vol. 6, no. 1, pp. 13–23, 2016, [Online]. Available: https://www.researchgate.net/publication/322200082

[32] W. Guo and Y.-J. Choi, “Assessing dimensionality of IRT models using traditional and revised parallel analyses,” *Educ Psychol Meas*, vol. 83, no. 3, pp. 609–629, Jun. 2023, doi: 10.1177/00131644221111838.

[33] J. Hattie, “Methodology review: Assessing unidimensionality of tests and itenls,” *Appl Psychol Meas*, vol. 9, no. 2, pp. 139–164, Jun. 1985, doi: 10.1177/014662168500900204.

[34] M. Gökcan and D. Çobanoğlu Aktan, “Validation of the vocabulary size test,” *Egit Psikol Olcme Deger Derg*, vol. 13, no. 4, pp. 305–327, Dec. 2022, doi: 10.21031/epod.1144808.

[35] E. Moradi, Z. Ghabanchi, and R. Pishghadam, “Reading comprehension test fairness across gender and mode of learning: Insights from IRT-based differential item functioning analysis,” *Language Testing in Asia*, vol. 12, no. 1, p. 39, Sep. 2022, doi: 10.1186/s40468-022-00192-3.

[36] H. Ahmad, N. Mamat, M. Che Mustafa, and S. Iryani Mohd Yusoff, “Validating the teaching, learning, and assessment quality of Malaysian ECCE instrument,” *International Journal of Evaluation and Research in Education (IJERE)*, vol. 10, no. 1, p. 135, Mar. 2021, doi: 10.11591/ijere.v10i1.20857.

[37] W. Astuti and Adiwijaya, “Support vector machine and principal component analysis for microarray data classification,” *J Phys Conf Ser*, vol. 971, p. 012003, Mar. 2018, doi: 10.1088/1742-6596/971/1/012003.

[38] S. T. Nihan, “Karl Pearsons chi-square tests,” *Educational Research and Reviews*, vol. 15, no. 9, pp. 575–580, Sep. 2020, doi: 10.5897/ERR2019.3817.

[39] W. J. van der Linden, Ed., *Handbook of item response theory*. Chapman and Hall/CRC, 2016. doi: 10.1201/9781315374512.

[40] A. Darmana, A. Sutiani, and Jasmidi, “Development of the thermochemistry- HOTS-tawheed multiple choice instrument,” *J Phys Conf Ser*, vol. 1462, no. 1, p. 012057, Mar. 2020, doi: 10.1088/1742-6596/1462/1/012057.

[41] W.-H. Chen and D. Thissen, “Local dependence indexes for item pairs using item response theory,” *Journal of Educational and Behavioral Statistics*, vol. 22, no. 3, pp. 265–289, Sep. 1997, doi: 10.3102/10769986022003265.

[42] D. O. Tobih, M. A. Ayanwale, O. A. Ajayi, and M. V. Bolaji, “The use of measurement frameworks to explore the qualities of test items,” *International Journal of Evaluation and Research in Education (IJERE)*, vol. 12, no. 2, p. 914, Jun. 2023, doi: 10.11591/ijere.v12i2.23747.

[43] C. D. Desjardins and O. Bulut, *Handbook of educational measurement and psychometrics using R*. Boca Raton, Florida : CRC Press, [2018]: Chapman and Hall/CRC, 2018. doi: 10.1201/b20498.

[44] S. Soeharto and B. Csapó, “Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts,” *Heliyon*, vol. 7, no. 11, p. e08352, Nov. 2021, doi: 10.1016/j.heliyon.2021.e08352.

[45] J. Jumini and H. Retnawati, “Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination,” *AL-ISHLAH: Jurnal Pendidikan*, vol. 14, no. 1, pp. 385–398, Mar. 2022, doi: 10.35445/alishlah.v14i1.926.

[46] H. Chin, C. M. Chew, W. Yew, and M. Musa, "Validating the cognitive diagnostic assessment and assessing students' mastery of 'parallel and perpendicular lines' using the rasch model," *Participatory Educational Research*, vol. 9, no. 6, pp. 436–452, Nov. 2022, doi: 10.17275/per.22.147.9.6.

[47] Herwin and S. C. Dahalan, "Person fit statistics to identify irrational response patterns for multiple-choice tests in learning evaluation," *Pegem Journal of Education and Instruction*, vol. 12, no. 4, pp. 39–46, Jan. 2022, doi: 10.47750/pegegog.12.04.05.