Hipotenusa : Journal of Mathematical Society
Volume 6 Issue 2 Year 2024
ISSN: 2716-3156
Website: https://ejournal.uinsalatiga.ac.id/index.php/hipotenusa/index

IRT Method for Measuring The Quality of High School Mathematics Mid-Semester Assessment Questions in Magelang

Gunawan[1]*, Kana Hidayati [2]
[1,2] Program Studi Magister Pendidikan Matematika, Universitas Negeri Yogyakarta, Indonesia
*Corresponding Author. E-mail: gunawan.2023@student.uny.ac.id[1]
kana@uny.ac.id[2]

DOI: 10.18326/hipotenusa.v6i2.2036

Article submitted: June 21, 2024
Article reviewed: November 29, 2024
Article published: December 30, 2024

Abstrak

Learning assessment is an important part of the learning process. The quality of the assessments carried out must of course be proven, namely by using item analysis, one of which is analysis using the IRT (Item Response Theory) method. This research aims to determine the quality of the question items that have been designed in mathematics mid-semester assessment questions using the IRT method. The method used in this research is descriptive analysis with a quantitative approach. The research subjects were 143 class XI students at one of the Magelang Regency high schools. The research instrument used was mathematics mid-semester assessment questions in Multiple Choice form, totaling 20 questions. Data analysis was carried out using the technique of analyzing the characteristics of the questions based on the IRT method, consisting of two stages, namely the assumption test and the analysis test. Based on the results of the analysis, it was found that the items met the assumption tests, namely the unidimensional, local independence, and invariance tests. Based on the model suitability test, a 2-PL model was obtained that was suitable for use, namely analyzing the different power parameters and the difficulty level of the questions. Based on the results of the estimation analysis on the question item parameters, it was concluded that 16 questions met the criteria for good quality questions and 4 questions that met the criteria for deficient questions.

Kata Kunci: assessment, irt methods, mathematics

INTRODUCTION

For years, students' abilities in mathematics have often been associated with difficulties in understanding the material and inappropriate learning strategies. Students usually find it difficult because mathematics material is complex and learning strategies may not suit their needs, so they have difficulty achieving understanding (Raj Acharya, 2017; Wilkinson, 2018; Ziegler & Loos, 2017). This is often associated with low student achievement in mathematics subjects. Apart from these factors, Awopeju & Afolabi

(2016) revealed that the nature of the test questions and the characteristics of students also play an important role in their mathematical abilities. Test questions that may be designed with a certain level of difficulty can influence how students can answer them. It is important to consider the role of assessment in learning because assessment not only functions as a tool for measuring student abilities but also as a means of identifying and overcoming obstacles in mathematics learning.

Learning assessment is an important component that cannot be separated from the entire learning process (Black & Wiliam, 2018; Memarian & Doleck, 2024). Assessment is simply defined as the process of determining what students know and can do (Looney et al., 2018). Assessments are often understood as questions or assignments in quizzes, tests, or other forms of evaluation. Assessment is an integral part of the learning process to improve the performance of teachers and students being assessed (Archer, 2017; Khairil & Mokshein, 2018). Assessment is used to measure the extent to which students have achieved the set learning targets (Baird et al., 2017; Barnes et al., 2017). Assessment measures the extent to which students have achieved the set learning targets. Assessment is the basis for measuring learning effectiveness, assisting teachers in developing appropriate learning strategies, and providing feedback to improve the quality of education (Mahlambi et al., 2022). Thus, assessment is not just about giving grades, but also about providing insight to enhance the overall learning process. The better the quality of the assessment, the easier it is for teachers to understand students' strengths and weaknesses (Gunawan & Asria, 2023).

Based on a recent review of more than 4,000 studies, it is proven that implementing effective assessment in the classroom can significantly increase student learning speed (Earl, 2013, p. 3). Well-designed and implemented assessments not only help students understand the material more quickly but also speed up their overall learning process. Therefore, teachers as educators need to have adequate knowledge and skills in planning assessments, observing the learning process, and providing constructive feedback to students (Gardner, 2012, p. 38). This is important to ensure that assessments function optimally in supporting student learning progress and increasing teaching effectiveness.

The quality of the assessment carried out by the teacher must of course be proven, namely by analyzing the question items (Halik et al., 2019; Hamimi et al., 2020). Analysis of question items is an important activity in creating questions to ensure the quality of the questions produced (Susanto et al., 2015; Wahiah et al., 2023; Yoshita Cahyaningrum et al., 2023). The need for question analysis arises because learning assessment is not only a process of providing grades but is also an important tool in improving overall learning effectiveness. Apart from that, according to (Aiken, 1994: 63), item analysis aims to improve the quality of the test by revising or deleting ineffective questions and obtaining diagnostic information about students' level of understanding of the material. Through this analysis, teachers can evaluate the extent to which previously designed learning objectives have been achieved. In addition, item analysis is important to improve the quality of items that will be reused in future tests (Quaigrain & Arhin, 2017). Item analysis can also be used to eliminate confusing or misleading question items in a test.

Methods for analyzing test items include two approaches, namely classical theory (classical test theory) and item response theory (item response theory) or the IRT method

(Kiliç et al., 2023). Over the last few decades, the IRT method has developed rapidly and has become an important complement to classical test theory in test development (Stage, 2003, p. 2). The IRT method emerged due to the limitations of classical test theory (Sainuddin, 2018). The IRT method was developed in the 1950s and 1960s by Frederic Lord and other psychometric experts. Their goal was to create a method that could evaluate respondents without relying on the same items in the test (Zanon et al., 2016). The IRT method has received much attention in instrument validation because it allows the estimation of students' abilities on any item (Gyamfi & Acquaye, 2023). The IRT method assumes that students' abilities are influenced by one dimension, namely the abilities measured in it, and the student's ability to answer one test item does not affect the answers to other items (Kong & Lai, 2022). In the IRT method, there are three parameters used, namely one parameter (1-PL), two parameters (2-PL), and three parameters (3-PL) models (Na et al., 2024). In the IRT model, three parameters can be used: one parameter (1PL), two parameter (2PL), and three parameter (3PL) models (Na et al., 2024). The 1-PL model only uses item difficulty level parameters, the 2-PL model uses difficulty level and difference power parameters, while the 3-PL model adds pseudo guessing parameters in addition to the item difficulty level and difference power (Bichi, 2015).

Treiblmaier et al. (2017) stated that there are several advantages when using the IRT method in analyzing test instruments. First, IRT assumes that the relationship between the answers on the test and the characteristics or abilities being measured is nonlinear. Second, by using IRT individual abilities can be performed more precisely, taking into account the unique characteristics of each test item. Third, IRT allows the estimation of item parameters (such as level of difficulty or discriminating power) independently of the sample used, so that the results are more general and can be applied to a wider population. Fourth, IRT not only allows the application of basic concepts such as reliability and internal consistency but also expands these concepts so that researchers can obtain more detailed information regarding the measurement process.

Based on this background, this research aims to analyze Mid-Semester Assessment questions in Mathematics using the IRT (Item Response Theory) method. This analysis aims to evaluate the quality of the question items that have been designed in the Mid-Semester Assessment. Using the IRT method, this research will measure the extent to which the question items in the Mathematics Mid-Semester Assessment can be considered of good quality.

METHODS

In this research, a descriptive analysis method with a quantitative approach was used, which aims to provide an overview of the results of the analysis of PTS Mathematics test items using the IRT (Item Response Theory) method. This research involved 143 class XI students at one of the high schools in Magelang Regency as research subjects. The instrument used is a PTS Mathematics question sheet in Multiple Choice (PG) form which consists of 20 questions, each question has 5 alternative answers (A, B, C, D, E). Data analysis was carried out using the technique of analyzing the characteristics of the questions based on the IRT method, consisting of two stages, namely assumption testing and test analysis. The assumption test consists of (1) unidimensional assumptions analyzed from the eigenvalues of the inter-item covariance matrix; (2) the assumption of local independence which is confirmed based on the test results on the unidimensional assumption; The assumption of parameter invariance was carried out by dividing respondents into two groups based on odd-even ordinal numbers.

In Item Response Theory (IRT), there are three main models used to analyze test data, namely the 1-PL, 2-PL, and 3-PL models. The 1-PL model, also known as the Rasch model, only takes into account one parameter, namely the level of item difficulty. This model assumes that all items have the same discriminating power so that only differences in level of difficulty are measured. The 2-PL model adds one more parameter, namely item discriminating power, which allows each item to have different abilities in differentiating between students with high and low abilities. The 3-PL model introduces a third parameter, namely the guessing parameter, which takes into account the probability of students answering correctly at random, especially in multiple-choice questions. The selection of the model used is very dependent on the results of the model fit test analysis, which is carried out to ensure the model fits the data being analyzed.

This research was carried out through several systematic stages. The first step is to prepare student answer data, which will be used in further analysis. After the data has been summarized, assumption testing is carried out which involves three important aspects, namely the unidimensional test to ensure that the data measures one main construct, the independence test to check whether students' answers to one item are not influenced by the answers to other items, and the invariance test to ensure that Item characteristics remained consistent across different groups of students. After these assumptions are met, this research continues with determining the model suitability test to select whether the 1-PL, 2-PL, or 3-PL IRT model best fits the student data. After the appropriate model is selected, item parameter analysis is carried out according to the model used. This analysis aims to evaluate the characteristics of the items in the test instrument, including how difficult the items are, how well the items differentiate between students with different levels of ability, and in the 3-PL model, how likely it is that students answer correctly at random.

RESULTS AND DISCUSSION

Test Assumptions

The first step before estimating test parameters is to test the assumptions underlying Item Response Theory (IRT). These assumptions include unidimensionality, meaning that the test measures one major construct or ability; local independence, which indicates that the response to each test item is not influenced by the response to other items after controlling for ability factors; and parameter invariance, which states that item parameters and ability parameters do not depend on certain subgroups of the population. Testing these assumptions is important to ensure that the IRT model used is appropriate and can provide accurate and valid parameter estimates.

Unidimensional Assumption

The unidimensionality assumption was checked by performing factor analysis, which included the eigenvalues of the inter-item covariance matrix. This analysis was carried out using SPSS software. The initial step in factor analysis is to assess sample adequacy using the KMO Test and Bartlett's Test, as shown in Table 1.

Table 1. KMO Test Results and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0,758
Bartlett's Test of Sphericity	Approx. Chi-Square	590,941
	df	190
	Sig.	,000

Based on Table 1, the KMO value is recorded at 0.758, while the Chi-Square value is 590.941 with degrees of freedom of 190 and a sig value. 0,000. Factor analysis can be fulfilled if $KMO > 0{,}5$ and $sig. < 0{,}05$ (Arlinwibowo et al., 2021). From these results, it can be seen that the total of 143 samples used in the research met the sample adequacy criteria needed to continue factor analysis. Factor analysis was then carried out using SPSS. The results of data processing for factor analysis can be found in the Eigen value section in Table 2 below.

Table 2. Eigen Value

Component	Total	Initial Eigenvalues % of Varians	Cumulative %
1	4.488	22.442	22.442
2	1.514	7.570	30.012
3	1.384	6.920	36.932
4	1.323	6.617	43.549
5	1.230	6.151	49.700
6	1.162	5.811	55.510
7	1.043	5.217	60.727
8	.981	4.906	65.633
9	.881	4.406	70.039
10	.835	4.174	74.213
11	.773	3.863	78.076
12	.709	3.546	81.622
13	.616	3.078	84.700
14	.542	2.712	87.413
15	.496	2.479	89.892
16	.487	2.434	92.325
17	.447	2.233	94.559
18	.414	2.069	96.627
19	.365	1.824	98.451
20	.310	1.549	100.000

Based on Table 2, 5 factors show an eigenvalue of more than 1. Based on this value, PTS has 7 factors. These 7 factors can explain 60.727% of the variance. These Eigen values can be visualized in the scree plot shown in Figure 1.

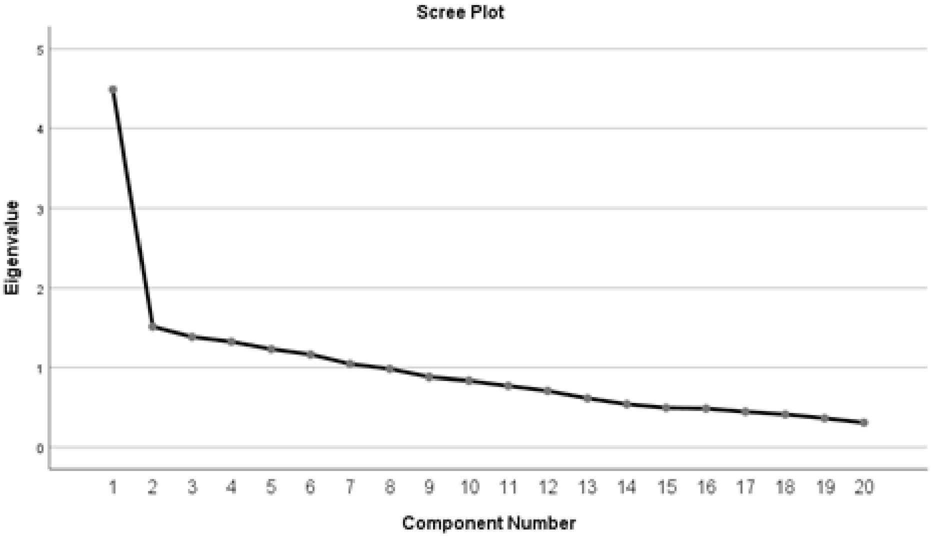

Figure 1. Output *Scree Plot*

From the illustration in Figure 1, there is a sharp decrease in the eigenvalue between factor 1 and factor 2. The eigenvalue begins to decrease at factor 2, so the scree plot almost forms a right angle. This shows that there is only one dominant factor in the test set, which indicates that the unidimensional assumption is met. Research conducted by Hartono et al. (2022) also confirmed these findings. They noted a sharp drop in eigenvalues between factor 1 and factor 2, supporting that the test had only one dominant factor. This confirms that unidimensional analysis can be applied in this test.

Assumption of Local Independence

The assumption of local independence is not tested separately in this context but is confirmed based on the results of previous unidimensional tests. This view is supported by Retnawati (2014, p. 7), who states that the assumption of local independence is automatically fulfilled if the response to the test is proven to be unidimensional. Therefore, because the unidimensionality assumption has been met, it can be concluded that the local independence assumption is also met. This is in line with research which states that local independence is fulfilled because the results of the unidimensional assumption are met (Özdemir, 2015).

Parameter Invariance Assumption

This assumption is substantiated through item parameter estimates. To test the invariance of item parameters, a 2-parameter (2-PL) model is used, which involves the level of difficulty and differentiability of the items. To test the invariance of item parameters, respondents were divided into two groups based on odd and even serial numbers. The results of the item parameter invariance analysis can be seen in Figure 2 and Figure 3 below.

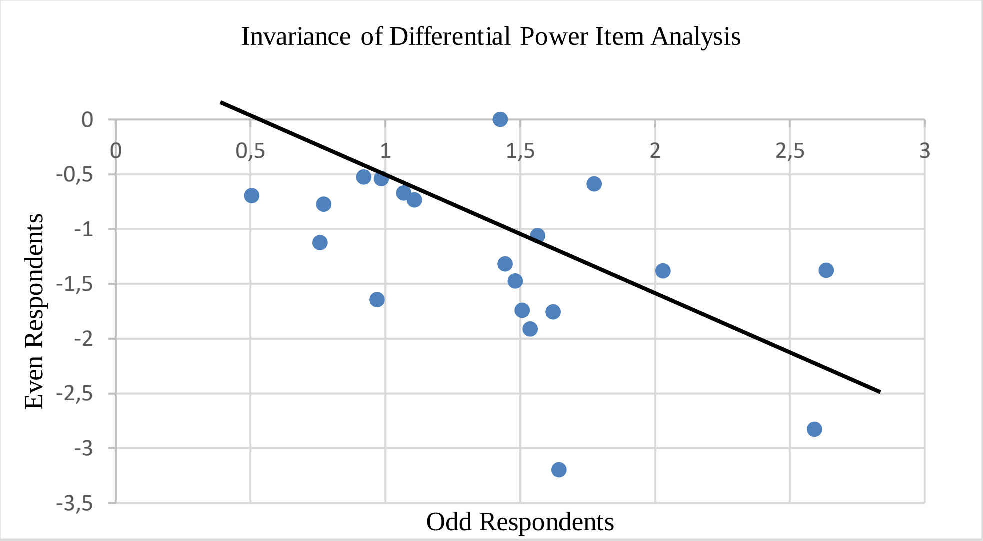

Figure 2. Invariance of Differential Power Item (a)

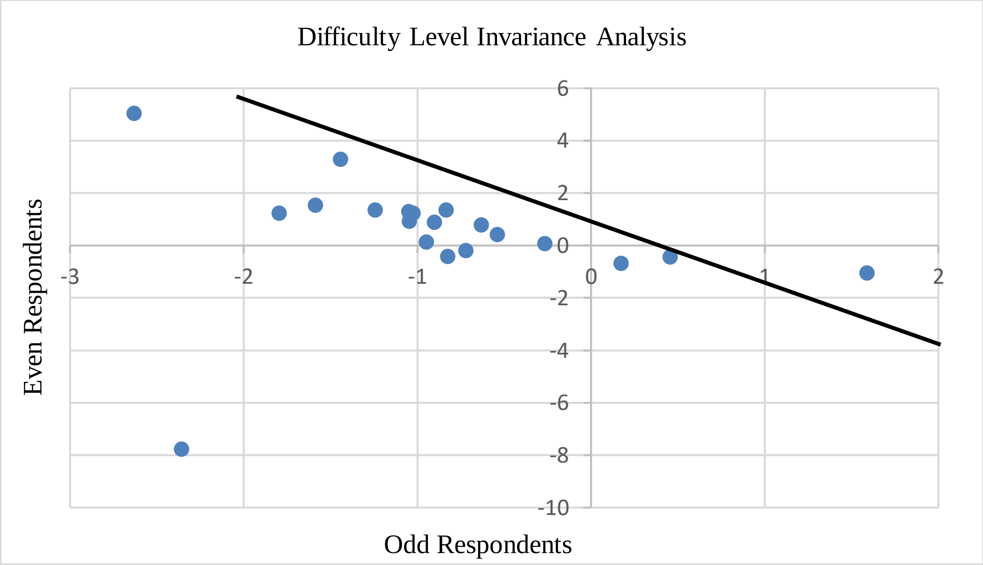

Figure 3. Difficulty Level Invariance (b)

Based on the results of the invariance analysis of the item parameters, each point is relatively close to the line $y = x$. This shows that there is no significant variation in the parameters of item differentiation (a) and item difficulty level (b) resulting from dividing respondents with odd and even serial numbers. Thus, it can be concluded that the invariance of item differentiation and item difficulty level has been fulfilled. This is in line with research by Apriyani et al. (2023) which stated that no violations were found of the assumption of parameter invariance or test-taker abilities.

Test Model Fit

3 models are tested for model suitability in PTS Mathematics questions, namely the 1-PL, 2-PL, and 3-PL models. The goodness-of-fit test is carried out through statistical goodness of fit based on the p-value. The question item is said to match the model if the p-value>α(α=0.05). Determining the model selected is based on the items that match the

most (Sumaryanta, 2021, p. 74). This model suitability test was analyzed using R software with the results shown in Table 3 below.

Table 3. Results of Model Fit Analysis

Question Number	1 Parameter (1-PL)		2 Parameter (2-PL)		3 Parameter (3-PL)	
	P-value	Decision	*P-value*	Decision	*P-value*	Decision
1	0,6866	Fit	0,2855	Fit	0,2337	Fit
2	0,0115	Misfit	0,1251	Fit	0,239	Fit
3	0,3941	Fit	0,4037	Fit	0,1672	Fit
4	0,2018	Fit	0,4033	Fit	0,1191	Fit
5	0,5491	Fit	0,7198	Fit	0,4090	Fit
6	0,6286	Fit	0,7238	Fit	0,1848	Fit
7	0,0053	Misfit	0,2114	Fit	0,5084	Fit
8	0,2118	Fit	0,1003	Fit	0,2045	Fit
9	0,0594	Fit	0,3781	Fit	0,2010	Fit
10	0,6753	Fit	0,4253	Fit	0,0817	Fit
11	0,6965	Fit	0,5652	Fit	0,1345	Fit
12	0,1348	Fit	0,4936	Fit	0,0005	Misfit
13	0,1105	Fit	0,3554	Fit	0,0193	Misfit
14	0,0031	Misfit	0,4209	Fit	0,1283	Fit
15	0,6913	Fit	0,7015	Fit	0,3877	Fit
16	0,5822	Fit	0,6874	Fit	0,1045	Fit
17	0,0479	Misfit	0,0019	Misfit	0,0442	Misfit
18	0,0480	Misfit	0,3330	Fit	0,2533	Fit
19	0,1060	Fit	0,0104	Misfit	0,2063	Fit
20	0,8787	Fit	0,3233	Fit	0,3243	Fit

Table 3 illustrates the suitability of the logistic parameter model for the 20 items. In the 1-PL model, there are 15 questions that match the model and 5 questions that do not. The 2-PL model shows that 18 items fit the model, while 2 items do not. Meanwhile, in the 3-PL model, 17 items match the model and 3 items do not match. Based on this analysis, the logistic parameter model chosen was 2-PL because it had the largest number of questions that fit the model, namely 18 questions.

Analysis of Question Item Parameter Estimation

In the two-parameter logistic model (2-PL), the two main parameters analyzed are difficulty level and discrimination. The 2-PL model is used in item response theory to model the probability that a student will answer an item correctly, influenced by these two parameters. The item parameter estimation was analyzed using R software, and the results are presented in Table 4.

Table 4. Analysis of the Quality of Model 2-PL Question Items

Question Number	Discrimination (a)	Difficulty Level (b)	Criteria
1	1,411	-2,917	Deficient
2	1,148	-1,882	Good
3	0,794	-3,316	Deficient
4	2,745	-1,275	Deficient
5	1,364	-1,009	Good
6	0,949	-0,136	Good
7	1,680	-0,879	Good
8	1,267	-0,444	Good
9	2,261	-0.448	Deficient
10	0,938	0,363	Good
11	0,894	-1,566	Good
12	1,908	-1,130	Good
13	1,717	-1,084	Good
14	1,252	-1,060	Good
15	1,581	-1,507	Good
16	0,562	-0,193	Good
17	1,439	-0,673	Good
18	1,700	0,447	Good
19	0,743	-0,456	Good
20	0,728	1,390	Good

The discrimination index (parameter a) is a parameter that indicates the ability of an item to differentiate between test takers with varying abilities, specifically between those with high and low abilities (Zanon et al., 2016). Items with a high discrimination index demonstrate a stronger ability to distinguish between different levels of ability. The higher the value of a, the more effective the item is at selecting students with varying levels of difficulty. Then, the difficulty level (parameter b) indicates the relative position of the item on the scale of the test takers' ability. The difficulty level parameter (b) typically ranges from negative to positive values (Ayanwale et al., 2022). A negative b value indicates that the item is relatively easy, as it is likely to be answered correctly by most test takers, while a positive b value indicates that the item is more difficult.

Next, to strengthen the item analysis presented in Table 4, a comparison with the Item Characteristic Curve (ICC) is needed. The ICC shows how the probability of a participant answering a question correctly varies according to their ability level (Stemler & Naples, 2021). In the context of the IRT model, ideally, this curve takes an S-shape (Sigmoid), reflecting the gradual relationship between a participant's ability and their likelihood of answering correctly (Cai et al., 2016). This S-shape indicates that the item has an appropriate level of difficulty, allowing it to effectively differentiate between low and high ability students. Furthermore, the curve's slope, which is not too steep, helps maintain a balance between the difficulty level and the item's discrimination power, ensuring that the item is neither too easy nor too difficult. This characteristic curve is generated from the output of the R software, which is displayed in Figure 4 below.

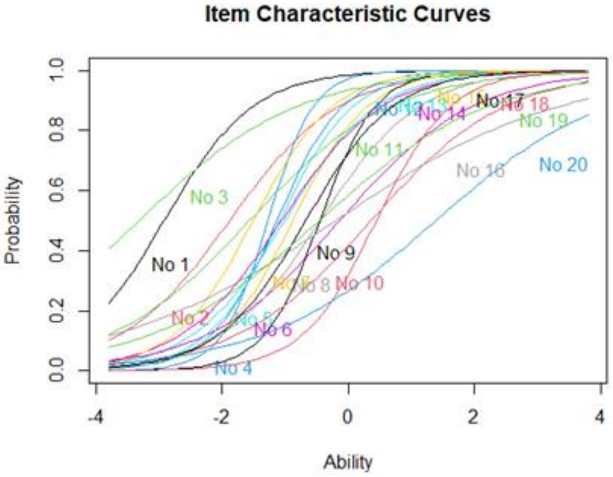

Figure 4. Item Characteristics Curve

Determination of question item quality criteria refers to Hartono et al. (2022), which determines that an item is considered good if it has a difficulty level in the range of -2.0 to 2.0 and a discriminating power between 0.0 and 2.0. The level of difficulty measures the extent to which the test taker can solve the question, where a score in this range indicates a question that is neither too easy nor too difficult. Discriminating power measures the ability of questions to differentiate between test takers who have high and low abilities.

Based on the results of the item analysis in Table 4, the quality of the 20 items analyzed is evaluated based on two parameters in the two-parameter logistic (2-PL) model: discrimination (a) and difficulty (b). From the table, 16 items, or 80%, meet the criteria for good quality. Items categorized as "good" have ideal values for discrimination and difficulty, making them effective in distinguishing between high and low-ability participants. Items with high discrimination are more effective in differentiating participants with varying abilities, while items with moderate difficulty levels indicate that they can be answered by participants across a range of ability levels.

Conversely, there are 4 items, or 20%, that do not meet these criteria, thus categorized as items with poor quality. This is because they do not satisfy one of the two parameters, namely discrimination and difficulty. According to Kusumayanti & Jannah (2022), items that do not meet these criteria may be too difficult or too easy, or have low discrimination power, making them less effective in accurately assessing the test participants' abilities. Such items need to be revised or improved before being included in the item bank to meet the expected quality standards (Retnawati & Hadi, 2014). Storing high-quality items in the item bank ensures more valid and reliable tests for future evaluations, while poor-quality items are recommended to be redeveloped or adjusted to improve their quality (Santoso et al., 2019).

CONCLUSION

Based on the IRT assumption test which includes unidimensional, local independence, and invariance assumptions, the results obtained are (1) the unidimensional assumption shows that there is only 1 dominant factor in the test set so that the unidimensional assumption is met, (2) the local independence assumption is also fulfilled because the local independence assumption is automatically proven after being proven by the unidimensionality of the response data to the test, (3) the assumption of parameter invariance with the 2-PL model that the invariance of item differentiation and item difficulty level is fulfilled. After the assumption test was met, a model suitability test was carried out which showed that 2-PL was the right model to use for this PTS Mathematics question. From the estimated values for each parameter, it was found that the different powers of the questions were in the range of 0.562 to 2.745. Then, the difficulty level of the questions is in the range of -3.316 to 1.390. Further analysis, namely reviewing the parameter values for differentiating power and level of difficulty, resulted in 16 questions meeting the criteria for good question quality and 4 questions meeting the criteria for deficient question quality.

REFERENCES

Aiken, L. R. (1994). Psychological testing and assessment (8th ed.). Allyn & Bacon.

Apriyani, D. C. N., Susanto, H. P., & Hidayat, T. (2023). Analysis of Pre-Olympic Middle School Mathematics Test Instruments Based on Item Response Theory. AlphaMath : Journal of Mathematics Education, 9(2), 145. https://doi.org/10.30595/alphamath.v9i2.18021

Archer, E. (2017). The Assessment Purpose Triangle: Balancing the Purposes of Educational Assessment. Frontiers in Education, 2(August), 1–7. https://doi.org/10.3389/feduc.2017.00041

Arlinwibowo, J., Achyani, I., & Kurniadi, G. (2021). Multidimentional Item Respose Utilization for Validating Mathematics National Examination in Indonesia. Journal of Physics: Conference Series, 1764(1). https://doi.org/10.1088/1742-6596/1764/1/012113

Ayanwale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The ftatus of the framework at the examination council of Lesotho. International Journal of Learning, Teaching and Educational Research, 21(8), 384–406. https://doi.org/10.26803/ijlter.21.8.22

Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: fields apart? Assessment in Education: Principles, Policy and Practice, 24(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Barnes, N., Fives, H., & Dacey, C. M. (2017). U.S. teachers’ conceptions of the purposes of assessment. Teaching and Teacher Education, 65, 107–116. https://doi.org/10.1016/j.tate.2017.02.017

Bichi, A. A. (2015). Comparison of Classical Test Theory and Item Response Theory: A Review of Empirical Studies Test Items Development and Analysis Using Item Response Theory View project. Australian Journal of Basic and Applied Sciences, 9(7), 549–556. https://doi.org/10.13140/RG.2.1.1561.5522

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. Assessment in Education: Principles, Policy and Practice, 25(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. Annual Review of Statistics and Its Application, 3, 297–321. https://doi.org/10.1146/annurev-statistics-041715-033702

Earl, L. M. (2013). Assessment as learning: Using classroom assessment to maximize student learning. Sage Publications.

Gardner, J. (2012). Assessment and learning. Sage Publications.

Gunawan, & Asria, L. (2023). Analisis butir soal Penilaian Tengah Semester (PTS) matematika kelas XI berdasarkan teori klasik. MATH LOCUS: Jurnal Riset Dan Inovasi Pendidikan Matematik, 4(1), 1–11.

Gyamfi, A., & Acquaye, R. (2023). Parameters and Models of Item Response Theory (IRT): A review of literature. Acta Educationis Generalis, 13(3), 68–78. https://doi.org/10.2478/atd-2023-0022

Halik, A. S., Mania, S., & Nur, F. (2019). Analisis Butir Soal Ujian Akhir Sekolah (Uas) Mata Pelajaran Matematika Pada Tahun Ajaran 2015/2016 Smp Negeri 36 Makassar. Al Asma : Journal of Islamic Education, 1(1), 11. https://doi.org/10.24252/asma.v1i1.11249

Hamimi, L., Zamharirah, R., & Rusydy, R. (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. Mathema: Jurnal Pendidikan Matematika, 2(1), 57. https://doi.org/10.33365/jm.v2i1.459

Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2022). Uji Kecocokan Model Parameter Logistik Soal Diagnosa Kemampuan Matematika Dasar. JNPM (Jurnal Nasional Pendidikan Matematika), 6(1), 125. https://doi.org/10.33603/jnpm.v6i1.5899

Khairil, L. F., & Mokshein, S. E. (2018). 21st Century Assessment: Online Assessment. International Journal of Academic Research in Business and Social Sciences, 8(1), 659–672. https://doi.org/10.6007/ijarbss/v8-i1/3838

Kiliç, A. F., Koyuncu, İ., & Uysal, İ. (2023). Scale Development Based on Item Response Theory : A Systematic Review. 10(1), 209–223.

Kong, S., & Lai, M. (2022). Computers & Education Validating a computational thinking concepts test for primary education using item response theory : An analysis of students ' responses. 187(May), 1–18.

Kusumayanti, A., & Jannah, N. (2022). Analisis butir soal ujian masuk mandiri UIN Alauddin Makassar dengan teori tes modern. MaPan: Jurnal Matematika Dan Pembelajaran, 10(1), 159–174.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2018). Reconceptualising the role of teachers as assessors: teacher assessment identity. Assessment in Education: Principles, Policy and Practice, 25(5), 442–467. https://doi.org/10.1080/0969594X.2016.1268090

Mahlambi, S. B., Studies, I., Africa, S., & Mahlambi, S. B. (2022). Mathematics teachers ' use of assessment for learning to promote classroom diversity of learners. Pythagoras-Journal of the Association for Mathematics Education of Africa, 1–9.

Memarian, B., & Doleck, T. (2024). Computers in Human Behavior : Artificial Humans A review of assessment for learning with artificial intelligence. Computers in Human Behavior: Artificial Humans, 2(1), 1–11. https://doi.org/10.1016/j.chbah.2023.100040

Na, C., Clarke-Midura, J., Shumway, J., van Dijk, W., & Lee, V. R. (2024). Validating a performance assessment of computational thinking for early childhood using item response theory. International Journal of Child-Computer Interaction, 40(May 2023), 100650. https://doi.org/10.1016/j.ijcci.2024.100650

O. A., A., & E. R. I., A. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. European Scientific Journal, ESJ, 12(28), 263. https://doi.org/10.19044/esj.2016.v12n28p263

Özdemir, B. (2015). A Comparison of IRT-based Methods for Examining Differential Item Functioning in TIMSS 2011 Mathematics Subtest. Procedia - Social and Behavioral Sciences, 174, 2075–2083. https://doi.org/10.1016/j.sbspro.2015.02.004

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. Cogent Education, 4(1), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Raj Acharya, B. (2017). Factors Affecting Difficulties in Learning Mathematics by Mathematics Learners. International Journal of Elementary Education, 6(2), 8. https://doi.org/10.11648/j.ijeedu.20170602.11

Retnawati, H. (2014). Teori respon butir dan penerapannya. Nuha Medika.

Retnawati, H., & Hadi, S. (2014). Sistem bank soal daerah terkalibrasi untuk menyongsong era desentralisasi. Jurnal Ilmu Pendidikan, 20(2), 183–193.

Sainuddin, S. (2018). Analisis Karakteristik Butir Tes Matematika Berdasarkan Teori Modern (Teori Respon Butir). Jurnal Penelitian Matematika Dan Pendidikan Matematika, 1(1), 1–12.

Santoso, A., Kartianom, K., & Kassymova, G. K. (2019). Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka). Jurnal Riset Pendidikan Matematika, 6(2), 165–176. https://doi.org/10.21831/jrpm.v6i2.28900

Stage, C. (2003). Classical test theory or item response theory: The Swedish experience. 42, 1–29.

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. item response theory: Knowing when to cross the line. Practical Assessment, Research and Evaluation, 26, 1–16. https://doi.org/10.7275/v2gd-4441

Sumaryanta. (2021). Teori klasik & teori respon butir: Konsep & contoh penerapannya. CV. Confident.

Susanto, H., Rinaldi, A., & Novalia. (2015). Analisis validitas reabilitas tingkat kesukaran dan daya beda pada butir soal Ujian Akhir Semester ganjil mata pelajaran matematika. 3(2), 203–217.

Treiblmaier, H., Rusch, T., Mair, P., & Lowry, P. B. (2016). Breaking Free from the Limitations of Classical Test Theory: Developing and Measuring Information Systems Scales Using Item Response Theory. Information & Management. https://doi.org/10.1016/j.im.2016.06.005

Wahiah, Z., Prabowo, S. M., & Safitri, H. A. (2023). Eksplorasi Efektivitas Tes Pilihan Ganda Berbasis Komputer Sebagai Evaluasi Pembelajaran. EDUCATIVO: Jurnal Pendidikan, 2(2), 342–347. https://doi.org/10.56248/educativo.v2i2.

Wilkinson, L. C. (2018). Teaching the language of mathematics: What the research tells us teachers need to know and do. Journal of Mathematical Behavior, 51(May), 167–174. https://doi.org/10.1016/j.jmathb.2018.05.001

Yoshita Cahyaningrum, I., Fuady, A., & Islam Malang, U. (2023). Analisis Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran Matematika Kelas VII dengan Berbantuan Aplikasi Software Anates. Mathema Journal, 5(2), 67–81.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. Psicologia: Reflexao e Critica, 29(1), 1–10. https://doi.org/10.1186/s41155-016-0040-x

Ziegler, G. M., & Loos, A. (2017). “What is Mathematics?” and why we should ask, where one should experience and learn that, and how to teach it. 63–77. https://doi.org/10.1007/978-3-319-62597-3_5