Optimizing Item Construction in Diagnostic Mathematics Test

Wahyu Hartono [1], Samsul Hadi [2], Raden Rosnawati [2]

[1] *Universitas Swadaya Gunung Jati/Doctoral Student at Universitas Negeri Yogyakarta,*
Jl. Pemuda No.32, Cirebon, Indonesia
[2] *Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta, Indonesia*

Abstract **–The diagnostic mathematics test is a critical tool for measuring students' abilities to understand and apply mathematical concepts, with the design of good test items being paramount to ensure validity. This study leverages Item Response Theory (IRT) models and Differential Item Functioning (DIF) methods to refine the construction of test items, specifically focusing on rational numbers. Engaging 929 junior high school students from three public schools in Cirebon, West Java The research utilized R Software to analyze the most suitable IRT models and investigate DIF methods. The findings underscore the efficacy of the Parameter Logistic 3PL model based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), -2 loglikelihood, and Standardized Root Mean Square Residual (SRMSR) values, alongside item fit, highlighting that nearly all analyzed items were suitable except one that required replacement. Additionally, the identification of items with significant DIF effects points to potential biases, suggesting avenues for enhancing test fairness and reliability. The study's broader implications extend to improving diagnostic assessment practices, informing item design in educational evaluations, and guiding future research towards creating more equitable and precise measures of mathematical understanding. This contributes to a nuanced comprehension of student abilities, offering valuable insights for educators, assessment designers, and policymakers aimed at fostering improved learning outcomes in mathematics education.**

Keywords **– Diagnostic mathematics test, DIF, IRT, rational number.**

DOI: 10.18421/TEM132-25
https://doi.org/10.18421/TEM132-25

Corresponding author: Samsul Hadi,
Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta, Indonesia
Email: samsul_hd@uny.ac.id

Received: 15 November 2023.
Revised: 13 February 2024.
Accepted: 22 February 2024.
Published: 28 May 2024.

1. Introduction

A mathematics diagnostic test is one of the evaluation tools commonly used to assess students' understanding and application of mathematical concepts. There are at least five stages in creating a test [1]: determine the test's objectives, identify the domain that needs testing, design the test framework, create a question bank, and validate the test. Ensuring test results accurately depict students' abilities is one element that needs consideration in diagnosing abilities. When an assessment contains the most negligible possible errors or no errors, it is considered accurate. The diagnostic test instrument must be valid to obtain accurate results demonstrating students' abilities, reliability, and appropriate item parameters.

We can estimate item parameters using Classical Test Theory (CTT) and Item Response Theory (IRT). In CTT, researchers examine the correlation coefficient between item scores and total test scores to determine the validity of test items. In contrast, in IRT, they calculate the validity of test items based on item discrimination and difficulty level [2]. IRT also considers respondents' characteristics in answering test items, while CTT does not consider respondents' characteristics in test analysis. The IRT models respondents' ability to master a concept or skill, allowing for high-accuracy estimation of individual proficiency [3]. On the other hand, CTT only produces relative test scores, indicating the comparison of respondents' abilities within the same test [2].

Both models have their respective advantages and disadvantages. CTT is simpler to use and easier to understand, making it commonly used in measurement practice. However, CTT pays less attention to item characteristics in detail and does not consider respondents' characteristics in answering test items. On the other hand, IRT is more complex in its application but can take into account item and respondent characteristics in detail, resulting in more accurate information about individual abilities [3].

Researchers regard IRT as a novel approach to addressing the shortcomings of classical test theory because it disentangles the relationship between items and the sample or subjects taking the test. The characteristics or abilities of test-takers remain the same regardless of the items. Conversely, item characteristics stay the same irrespective of the test-taker's abilities. Additionally, item response theory analyzes individual items rather than the entire test.

According to Stone [4], the model's failure to fit the data can cause problems with item parameter estimation. In IRT, Hambleton *et al.* [5] explained three logistic models: the one-parameter logistic (1PL), 2PL, and 3PL. The 1PL model has only one difficulty parameter (b). This parameter influences or determines the test-takers' characteristics (ability). In other words, the difficulty parameter of the item can be used to measure or determine a test-taker's ability. Test-takers with high abilities find it easy to answer the test items, while those with low abilities struggle to answer them. Items are good if their difficulty levels range from -2 (easy) to +2 (difficult).

The item difficulty parameter (b), the item discrimination (a), and the guessing (c) make up the three-parameter logistic model (3PL). A guessing parameter (c) represents the likelihood that a low-ability test-taker will accurately guess their response to a challenging issue. The 3PL model's guessing factor calculates the likelihood that respondents will accurately respond to a question merely by guessing or failing to comprehend the question. The 3PL model's guessing factor helps detect elements that could lead to guessing and, perhaps, lower test validity [6]. The value of a guessing parameter (c) is between 0 and 1. An item is considered good if the parameter "c" is less than 1/k, where "k" specifies the number of response alternatives.

On the other hand, we can observe the graphical method from the item characteristic curve. Using the curve, we evaluate the data distribution's correctness in relation to the model. The model is adequate if there is very little distance between the points and the match line [7]. The final determination of the proper or fit parameter model is the same as the statistical method, based on most items that fit the logistic parameter types (1PL, 2PL, and 3PL).

Obtaining accurate and unbiased measurement results when using a model to assess a construct relies on ensuring the validity and reliability. In this context, bias is a systematic error in the measurement process. Therefore, researchers should base the final determination of the appropriate parameter model on the majority of fit items with the logistic parameters (1PL, 2PL, and 3PL), and they should also use graphical methods such as the ICC to assess how accurately the data distribution compares to the model. Thus, using valid and reliable models will help reduce bias in the test and ensure accurate measurement results. In addition to bias caused by the inappropriate selection of IRT models (1PL, 2PL, or 3PL), bias in test items can also arise from other conditions, such as gender differences, regional differences, race, and other factors. This situation is namely as differential item functioning (DIF).

Members of two demographic groups with the same skill level may perform differently on the same item, known as differential item functioning (DIF). DIF can be characterized as variations in the item response function between groups according to the IRT's viewpoint. A test question in mathematics using sports-related phrases that men are more likely to understand than women is a prime example of DIF. Researchers expect items like this to exhibit DIF towards females, meaning they are less likely to yield correct responses than males with equivalent mathematical abilities. However, in reality, the causes of DIF are often more ambiguous [8].

The phrase "differential item functioning" (DIF) has primarily taken the role of the phrase "item bias" in the literature on IRT. DIF happens when two or more groups of test takers do not have the same association between a test item and the latent variable (or multidimensional latent vector). The relationship between the item response and the latent variable must differ between groups for an item to display DIF and any other item attribute [3]. Educators and researchers should avoid biased items in educational assessment as they can favor or disadvantage certain groups. In this study, the detection of DIF is conducted using IRT 3PL.

Researchers define DIF as the difference between the focal group and the reference group's likelihood of correctly answering an item in unidimensional item response theory. Since "how big the difference" is between the two groups indicates the amount of DIF displayed on the characteristic curve.

The R software will calculate DIF using the Raju index in real-world applications. If the Raju statistic value is less than -1.96 or greater than 1.96, researchers at a significance level 0.05 label an item as having differential item functioning (DIF). Similarly, a p-value of less than 0.05 might be used to identify it [9].

Based on the preceding explanation, this study seeks to use an item response theory (IRT) approach to examine the instrument's quality and determine the functioning of differential items (DIF) in the diagnostic test of mathematical abilities, focusing on the topics of rational numbers.

The primary objective of this work is to center on enhancing and streamlining a diagnostic mathematics assessment by employing sophisticated statistical techniques, particularly in relation to rational numbers. This research aims to provide significant contributions to the domain of educational measurement.

The research study's novelty is attributed to its extensive utilization of sophisticated statistical techniques, its concentration on a particular mathematical concept (rational numbers), the incorporation of a significant sample size, and the practical implications derived from the analysis, such as item replacement and insights into differential item functioning (DIF) effects.

2. Methodology

This study uses a quantitative approach to descriptive research. Item response theory conveys the item quality in the mathematical ability diagnostic instrument, specifically in the topic of rational numbers. The replies of junior high school students in the seventh grade to a 25-item diagnostic test yielded dichotomous results. The research subjects are 929 junior high school students from three public schools in Cirebon, West Java, Indonesia. The diagnostic test consists of 25 multiple-choice items with four options, developed based on the essential competencies of rational number. The analysis will include testing IRT assumptions, determining the best-fitting IRT model, estimating item parameters, and determining the information function. Gender analyses of differential item functioning (DIF) will be centered on 929 students, with 433 male students serving as the reference group and 496 female students serving as the focal group. In this study, researchers will identify DIF using the Raju index method. Researchers will conduct all analyses using the R program.

3. Results

The Bartlett test measures data homogeneity, whereas the KMO-MSA test evaluates sample suitability. If the Kaiser Meyer Olkin (KMO)-MSA value is more than 0.5, and the significance of the Bartlett test is lower than 0.05, factor analysis can continue.

We acquired the KMO-SMA and Bartlett values based on response data from this investigation, shown in Table 1.

Table 1. KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	**Bartlett's Test of Sphericity**		
	Approx. Chi-Square	**df**	**Sig.**
,862	4362,598	300	,000

Based on Table 1, the KMO-MSA value is 0.862 (interpreted as meritorious), and the significance level for Bartlett's test is 0.000. It shows that the number of sample satisfies criteria and it is homogeneous, so researcher can perform factor analysis [10]. The eigenvalues portion of Table 2 in Statistical Package for the Social Sciences (SPSS) displays the factor analysis results.

Table 2. Eigenvalue

Component	***Initial Eigenvalues***		
	Total	***% of Varians***	***Cumulative %***
1	5,027	20,107	20,107
2	1,978	7,911	28,017
3	1,648	6,594	34,611
4	1,333	5,331	39,942
5	1,225	4,901	44,844
6	1,008	4,034	48,878

Eigenvalues larger than one, according to Table 2, denote a single factor. The entire test instrument contains six elements based on these eigenvalues. These six variables can explain 48.878% of the variance. Researchers can then display these eigenvalues in Figure 1 as a scree plot.

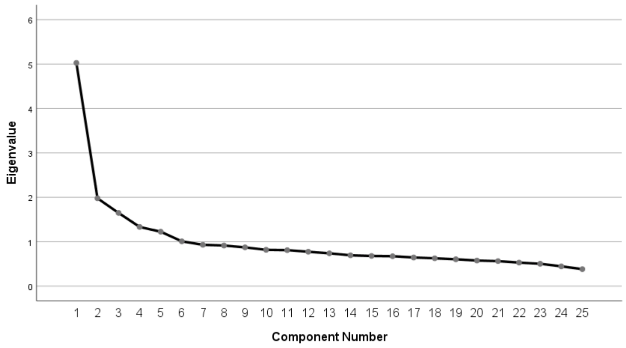

Figure 1. Scree plot factor analysis

The eigenvalues start to slope from Factor 3, almost creating a right angle, and the scree plot reveals a dramatic fall between Factor 1 and 2. These findings show that the test instrument only has one dominant element that satisfies the unidimensional assumption.

Local independence is the following presumption, which states that the response of a subject to one item has no influence on their response to next items. Local independence is satisfied when constant performance factors indicate that a subject's reaction to any item pair will be statistically independent [11].

Table 3 presents the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), -2 loglikelihood, and Standardized Root Mean Square Residual (SRMSR) values for each logistic parameter (1PL, 2PL, and 3PL). Referring to these values, it is apparent that IRT 3PL has the smallest value [12]. Therefore, the IRT 3PL model is considered the most suitable for analysis.

Table 3. Model selection

Model	AIC	BIC	-2 loglik	SRMSR
1 PL	23936.44	24178.15	23911.75	0.0695
2 PL	23938.27	24184.81	23836.27	0.0679
3 PL	23720.85	23851.37	23666.85	0.0605

Table 4 shows the p.S_X2 values for each item in the 1PL, 2PL, and 3PL models. The items composing the instrument best fit the 3PL model (3 Parameter Logistic Model). Therefore, we choose the IRT 3PL model for further analysis.

Table 4. Fit (+) and not fit (-) of data to IRT models 1PL, 2PL, and 3PL for each item

Item	1 PL		2 PL		3 PL	
1	0.61	+	0.58	+	0.66	+
2	0.04	-	0.83	+	0.62	+
3	0.00	-	0.00	-	0.00	-
4	0.08	+	0.05	+	0.05	+
5	0.03	-	0.08	+	0.10	+
6	0.68	+	0.97	+	0.52	+
7	0.01	-	0.03	-	0.04	-
8	0.11	+	0.01	-	0.04	-
9	0.00	-	0.00	-	0.00	-
10	0.23	+	0.06	+	0.04	-
11	0.00	-	0.09	+	0.28	+
12	0.00	-	0.70	+	0.81	+
13	0.10	+	0.09	+	0.05	+
14	0.12	+	0.53	+	0.57	+
15	0.07	+	0.17	+	0.08	+
16	0.00	-	0.11	+	0.06	+
17	0.13	+	0.16	+	0.14	+
18	0.03	-	0.09	+	0.19	+
19	0.03	-	0.08	+	0.05	+
20	0.22	+	0.37	+	0.05	+
21	0.00	-	0.00	-	0.00	-
22	0.01	-	0.01	-	0.06	+
23	0.00	-	0.00	-	0.07	+
24	0.00	-	0.01	-	0.06	+
25	0.00	-	0.00	-	0.02	-
Total Item Fit	10		16		18	

Further analysis involves determining the item parameters using the 3-PL model, which includes discrimination (a), difficulty (b), and pseudo-guessing (c) factors.

Based on the data in Table 5, the item difficulty ranges from -16.278 to 1.386; discrimination ranges from -0.103 to 5.099, and pseudo-guessing ranges from 0.001 to 0.315.

Table 5. Item parameter for discrimination (a), difficulty (b), and guessing (c)

Item	*a*	*b*	*c*	Evidence
1	0.823	-3.517	0.032	Not good (b < -2.000)
2	0.779	-1.122	0.002	good
3	1.153	-1.951	0.002	good
4	1.309	-2.226	0.003	Not good (b < -2.000)
5	1.311	1.178	0.062	good
6	1.69	-1.18	0.001	good
7	1.892	0.368	0.034	good
8	1.688	0.385	0.037	good
9	2.013	0.348	0.006	good
10	1.947	0.071	0.106	good
11	2.917	1.305	0.178	good
12	2.431	0.224	0.033	good
13	1.355	0.467	0.101	good
14	0.701	-2.564	0.019	Not good (b < -2.000)
15	1.624	0.526	0.098	good
16	1.39	1.311	0.315	Not good (c > 0.25)
17	3.299	0.832	0.286	Not good (c > 0.25)
18	1.529	1.373	0.117	good
19	1.155	-0.944	0.005	good
20	1.137	0.82	0.059	good
21	-0.103	-16.278	0.023	Not good (a < 0.00 and b < -2.000)
22	4.355	1.186	0.117	Not good (a > 2.000)
23	5.099	1.351	0.081	Not good (a > 2.000)
24	3.155	1.119	0.092	Not good (a > 2.000)
25	4.469	1.386	0.149	Not good (a > 2.000)

Analysis of the data indicates that out of the 25 items included in the diagnostic test instrument, only 18 items are appropriately aligned with the model's specifications, indicating a suitable fit. Furthermore, among these, 15 items are identified as being of high quality based on the established evaluation criteria. Although we can revise some items, we should discard or replace others, such as item 21, with alternative items. Item number 21 should be replaced because it has a negative discrimination parameter value and a significant negative difficulty value. Figure 2 shows the ICC curve for each item. Consistent with the previous analysis, we can see that overall, each item has a good item characteristic curve (ICC), except for item number 21.

Other than item 21, the form of the curve resembles a slightly tilted S, indicating that the likelihood of correctly answering an item rises as students' abilities rise.

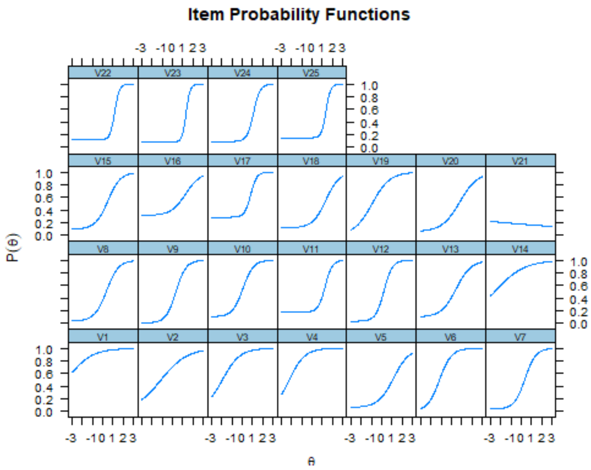

Figure 2. Item characteristic curve (ICC) for each item

The information function can also be analyzed using IRT analysis. The information function is a way to justify the selection of test items, the strength of a test item, and the evaluation of various test instruments [7].

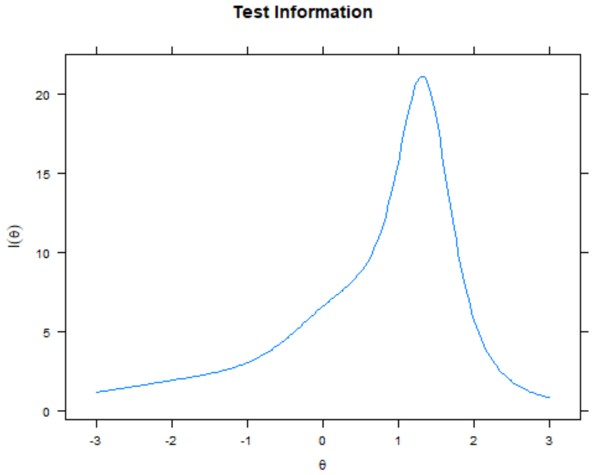

Figure 3. Test instrument information function

The test gives most information for students with logit abilities, around 1, based on the information function in Figure 3. The test instrument provides the most helpful information for students with above-average abilities. However, considering its relatively even distribution, this test instrument can also be considered reliable for measuring students' abilities within a logit range of -3 to +3. The analysis of item parameters resulted in removing item 21 out of 25 test items, leaving 24 items. Analysis continued with detecting differential item functioning (DIF) based on gender. DIF identification using the Raju area index approach is shown in Table 6.

Table 6. Statistic values and p-values using the Raju's area index method

Item	Stat.	P-value
1	-0.113	0.910
2	1.769	0.077
3	0.761	0.447
4	0.382	0.703
5	-0.166	0.868
6	1.338	0.181
7	3.227	0.001 *
8	4.055	0.000 *
9	1.622	0.105
10	2.009	0.045*
11	-1.157	0.247
12	1.889	0.059
13	-1.103	0.270
14	-1.171	0.242
15	2.238	0.025*
16	-0.928	0.353
17	-0.683	0.495
18	1.913	0.056
19	3.166	0.002*
20	-0.895	0.371
22	-0.978	0.328
23	0.941	0.347
24	-0.251	0.802
25	-1.356	0.175

'*'; Detection thresholds: -1.96 and 1.96 (significance level: 0.05);
Items detected as DIF items: 7, 8, 10, 15, and 19

Based on Table 6, we can see that five of the 24 test items have differential item functioning (DIF). These items are 7, 8, 10, 15, and 19. We identified these five items as having DIF because their stat. values fall outside the range of -1.96 to 1.96, or their stat. values are less than -1.96 or greater than +1.96, as presented in Figure 4. Figure 4 shows five items (7, 8, 10, 15, and 19) colored in red over the threshold value, denoting the presence of DIF. It is clear from these items that item 8 deviates significantly from the critical value, whereas item 10 deviates marginally from that value. It shows that item 8 has the most significant DIF effect, and item 10 has the most minor DIF effect. As shown in Figures 5 to 9, the difference in the areas under the ICC curves for the two groups can also be used to determine the presence of DIF.

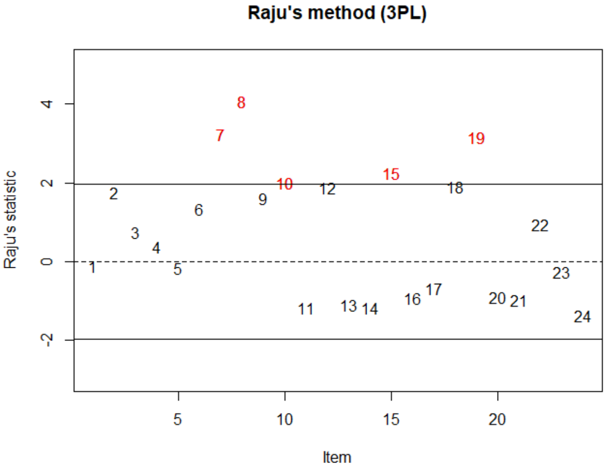

Figure 4. Item statistics values of the test instrument

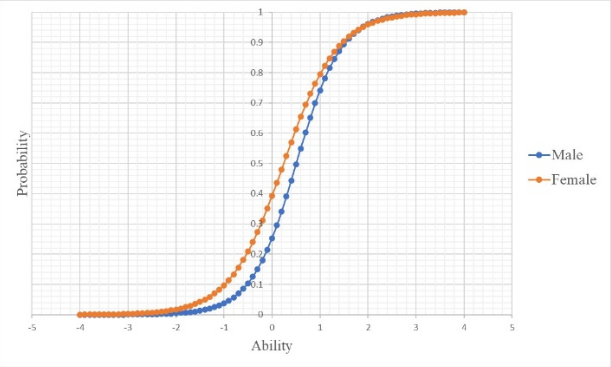

Figure 5. The difference in the area of the ICC curve for item number 7 between the two groups

In Figure 5, item 7 is more favorable towards females. It implies that women are more likely than men to provide accurate answers. It is evident from the difference in ICC curves, where the ICC curve for females is above the ICC curve for males within the ability range of -4 to 4.

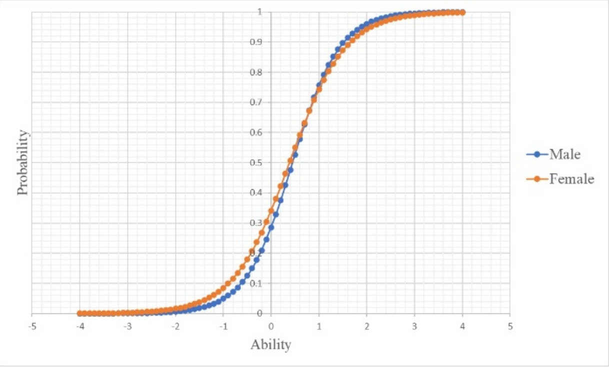

Figure 6. The difference in the area of the ICC curve for item number 8 between the two groups

In Figure 6, item 8 favors females with abilities between -2 and 1 more than males and favors males with abilities between 1 and 4 more than females. The two ICC curves intersect at an ability level of around 1.

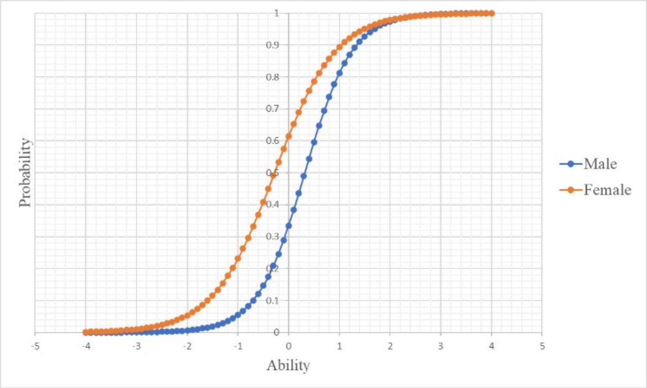

Figure 7. The difference in the area of the ICC curve for item number 10 between the two groups

In Figure 7, item 10 is more favorable towards females. It means that females are more likely to answer correctly than males. From the difference in ICC curves, we can observe that the ICC curve for females is above the curve for males in the ability range from -4 to 4.

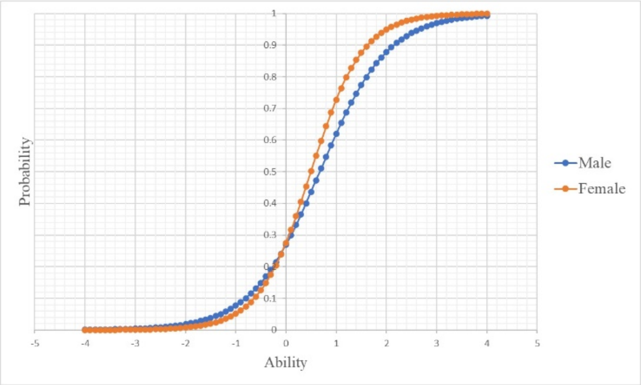

Figure 8. The difference in the area of the ICC curve for item number 15 between the two groups

In Figure 8, item 15 is more favorable towards females for abilities between 0 to 4 and more favorable on males for abilities below 0. The two ICC curves intersect around an ability of 0.

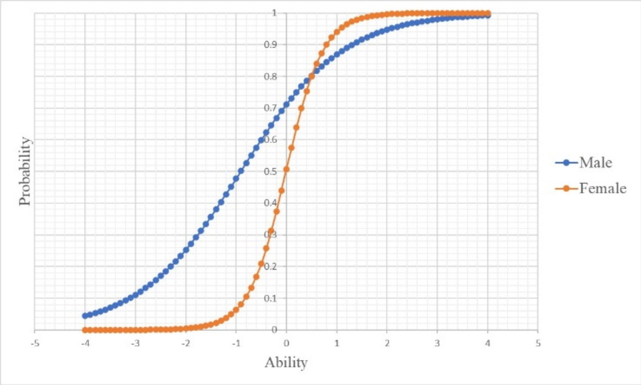

Figure 9. The difference in the area of the ICC curve for item number 19 between the two groups

In Figure 9, item 19 is more favorable towards females for abilities between 0.5 to 4 and more favorable on males for the abilities below 0.5. The two ICC curves intersect around an ability of 0.5.

4. Discussion

The fundamental assumptions in IRT analysis consist of three aspects: unidimensionality, local independence, and parameter invariance. Each item only measures one ability according to the concept of unidimensionality [7]. The application of factor analysis is justified by rigorous examination, which includes conducting feasibility tests such as the KMO-MSA and Bartlett tests. These tests ensure the suitability of the sample and the homogeneity of the data. The eigenvalues that have been acquired, especially those that are more than one, indicate significant factors. The scree plot demonstrates a clear decrease in magnitude from the first factor to the second factor, and a noticeable change in the rate of decline at the third factor. This supports the claim of unidimensionality. The identification of a prominent factor in the assessment tool confirms its alignment with the unidimensional assumption, establishing a strong basis for further examination and underscoring the accuracy of the diagnostic mathematics test in evaluating a single underlying skill.

The following assumption to consider is local independence. According to [2], researchers can assess local independence by confirming the unidimensional assumption. The concept of local independence within the framework of IRT analysis is based on the establishment of a logical connection with the previously verified premise of unidimensionality. By establishing the unidimensionality of the variables, we automatically validate the assumption of local independence, so strengthening the reliability of the selected statistical model. Transitioning to a different aspect of the discourse, the conversation smoothly advances towards the crucial undertaking of selecting an appropriate model. By utilizing the Akaike Information Criterion (AIC), BIC, -2 loglikelihood, and SRMSR values as a rigorous measure, the research effectively finds the IRT 3PL model as the most appropriate choice, based on its lowest value which signifies a stronger alignment between the model and the data [13]. Furthermore, a comprehensive examination of item fit, as shown by the p.S_X2 values, enhances the complexity of the process of selecting a model. The significant occurrence of compatible matches inside the 3PL model highlights its suitability for subsequent data analysis [12].

The utilization of a dual-method approach in this study serves to establish a strong statistical basis and emphasizes the importance of evaluating both the overall fit of the model and the performance of individual items during the selection process. Consequently, this approach strengthens the justification for employing the IRT 3PL model in the specific analysis being conducted.

In the study that follows, the 3-parameter logistic (3PL) model—which includes discrimination (a), difficulty (b), and guessing parameter (c)—is used to determine the item parameters. An item must fulfill specific requirements in order to be good [14], [15]. First, the item's difficulty level (b) should fall from -2 to +2. This range ensures the item is neither easy nor difficult for the test-takers. Second, the discrimination parameter (a) should be between 0 and 2. This range shows that the item successfully separates people with high and low abilities. Third, the guessing parameter (c) should not be greater than 0.25 or (1 / k), where k is the item's total number of response options. This criterion ensures that the item does not rely excessively on guessing and provides meaningful responses even for test-takers who are uncertain about the correct answer. By examining these item parameters and ensuring they meet the specified criteria, researchers can identify good-quality items for inclusion in the subsequent data analysis.

Based on the examination of item parameters, it is evident that of the 25 items comprising the diagnostic test instrument, only 19 items are deemed suitable for the data. Based on this evaluation, we observe that the items in the instrument predominantly fit the 3PL model, leading us to select the IRT 3PL model as the most appropriate model for subsequent data analysis. Typically, items with discrimination values greater than 2 are considered less desirable. If items exceed this threshold, researchers should try to identify the reasons behind their performance. Potential improvements include revising ambiguous items to avoid confusing high-ability students in their responses. The objective is to ensure that items can effectively differentiate between students who understand the material and those who do not, creating a reliable diagnostic test instrument. Item 21 is particularly problematic, given its negative discrimination parameter value and a highly negative difficulty value, indicating it is an easy item. A negative discrimination value implies that item 21 fails to distinguish between students with high and low abilities. That leads to the possibility of high-ability students answering incorrectly while low-ability students answer correctly. To maintain the diagnostic test instrument's integrity and alignment with its purpose, we should replace item 21 with another item.

Additionally, the item probability function (IPF), often called the item characteristic curve (ICC), can be used to evaluate the instrument's quality. This curve shows that respondents' likelihood of correctly responding to an item improves as skill levels rise. Consistent with the prior analysis, we observe that each item exhibits a favorable ICC with a curve shape closely resembling a tilted letter S, except for item 21. The trend suggests that as students' abilities improve, so does their likelihood of successfully answering a question.

IRT analysis also includes the examination of the information function, which serves as a valuable tool for understanding the strength of individual items within a test, aiding in the selection of test items, and facilitating comparisons between multiple test instruments [7]. The standard error of measurement (SEM), also referred to as the measurement error in IRT, has an inverse quadratic connection with information function. It suggests that smaller SEM values are correlated with higher values of information function. In contrast, lower values of information function associate with larger SEM values [2]. Information function helps to assess the reliability of the constructed test instrument; the more significant the information obtained from the function, the smaller the SEM will be. This relationship underscores the significance of information function in providing valuable insights into the precision and accuracy of the test instrument's measurements.

The test is most informative for students with a logit ability of around 1, showing that it delivers the most accurate measurement for those with above-average mathematical abilities. Nevertheless, due to its relatively even distribution, the test instrument can be considered reliable for assessing students' mathematics abilities across a broader range of logit abilities, from -3 to +3. The test demonstrates sufficient reliability in evaluating students' proficiency levels, encompassing those with meager abilities and those with very high abilities.

Subsequently, after scrutinizing the item parameters, a decision was made to exclude item 21 from the original set of 25 test items, leaving 24 items remaining for further analysis. The next step in the assessment involves detecting DIF based on gender, employing Raju's area index method. In order to get meaningful insights into potential biases or inequalities in the performance of the test instrument depending on gender, this DIF study seeks to determine whether there are any appreciable discrepancies in item responses between male and female students. DIF pertains to a scenario involving a single-point item, in which there is only one correct answer.

In this situation, the responding population is divided into two groups commonly referred to as the reference group and the focal group [17].

It is crucial to ensure that tests are fair for students with different characteristics, such as race, social background, and gender. Due to the potential biases introduced by items with DIF, which can undermine the validity of diagnostic evaluations [18], [19]. Researchers will incorporate the outcomes of analysis concerning the quality of item parameters and the identification of items showing DIF into further research on computer-adaptive testing (CAT)-based diagnostic assessments. CAT-based diagnostic assessments heavily rely on information related to items indicating DIF. The importance of DIF identification in CAT-based assessments over non-adaptive test formats is increased for some reasons. First, the CAT has fewer items, and the weight given to each response in calculating participants' test scores is more significant than a non-adaptive test.

Consequently, any deficiencies in an item can lead to more pronounced consequences. Furthermore, CAT present the item sequence to test takers is partly influenced by their responses to previous items, particularly those showing deficiencies. This aspect of the CAT process is where item deficiencies can considerably impact [8]. Therefore, detecting and addressing DIF in CAT-based diagnostic assessments becomes critical to maintaining fairness, validity, and accuracy in evaluating students' abilities and characteristics.

Differential item functioning (DIF) is present in five of the 24 items that made up the diagnostic test. Item 7, Item 8, Item 10, Item 15, and Item 19 are affected by the DIF. Upon analysis, it becomes evident that item 8 shows a significant deviation from the critical value, indicating the most considerable DIF effect among all the identified items. On the other hand, item 10 demonstrates a slight deviation from the critical value, signifying the lowest DIF effect, as described by Eren *et al.* [19]. Those five items are writes in the form of mathematical expressions, terms, and equations (not word problem). This finding significantly diverges from the results reported by Kan and Bulut [16]. In their study, certain word problems were identified as showing gender-related DIF favoring female examinees, whereas items expressed in mathematical terms showed similar performance across male and female examinees. Conversely, our analysis, which includes assessing DIF by examining the disparities in the item characteristic curve (ICC) between the two groups – specifically between male (reference group) and female students (focal group) – sheds light on potential biases or variances in item responses based on gender, suggesting a different pattern in item performance.

5. Conclusion

The diagnostic math test was examined using the IRT model, which produced some significant discoveries. First, based on the AIC, BIC, -2 loglikelihood, SRMSR values and the items that fit the model, the 3-parameter logistic (3PL) was the most appropriate IRT model. Secondly, we deemed the overall parameter estimation of the items in the instrument excellent, as 19 out of the 25 items exhibited suitable parameters. Thirdly, examining the item characteristic curves (ICC) for the 25 analyzed items revealed that 24 demonstrated favorable characteristics. In contrast, we should replace one item (item number 21) due to its inadequacy. Fourthly, the information function analysis indicated that the test instrument provided high levels of information for students with above-average abilities (around logit +1). Nonetheless, the test was still considered reliable for measuring the abilities of a diverse range of students, spanning from low to high abilities (ranging from logit -3 to logit +3).

Researchers use Raju's area index method and IRT model to determine differential item functioning (DIF) as part of the analysis to evaluate the effectiveness of the diagnostic test for mathematical ability. Among the 24 items that comprise the diagnostic test instrument, researchers have identified five items that exhibit DIF: items 7, 8, 10, 15, and 19. To improve the test instrument, researchers should direct efforts toward revising these five items. Potential areas for improvement include refining the wording of the items or adjusting the numerical values used within them. Addressing the differential item functioning (DIF) in these items can enhance the diagnostic test instrument's fairness and accuracy in measuring mathematics ability.

Acknowledgements

This study was a part of a dissertation. Thanks to the Government of Indonesia for funding this research through the Directorate of Resources, Directorate General of Education, Ministry of Education, Research and Technology. Contract of 2023 Number: 042/E5/PG.02.00.PL/2023

References:

[1]. Haladyna, T. M., & Downing, S. M. (2006). *Handbook of Test Development*. Lawrence Erlbaum Associates, Publishers.

[2]. DeMars, C. (2010). *Item Response Theory*. New York: Oxford University Press.

[3]. Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Psychology Press.

[4]. Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331-352.

[5]. Hambleton, R. K., Swaminathan, H., Rogers, D. J. (1991). *Fundamentals of Item Response Theory.* Library of Congress Cataloging-in-Publication Data. SAGE Publications.

[6]. Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

[7]. Retnawati, H. (2014). Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana (Item Response Theory and its Applications: For researchers, measurement and testing practitioners, and postgraduate students). *Nuha Medika*.

[8]. Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Springer Science & Business Media.

[9]. Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207.
Doi: 10.1177/014662169001400208

[10]. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. (2018). *Multivariate Data Analysis.* Cengage India.

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

[12]. Lee, Y., Rojas-Perilla, N., Runge, M. et al. (2023). Variable selection using conditional AIC for linear mixed models with data-driven transformations. *Statistics and Computing*, *33*, 1-17.
Doi: 10.1007/s11222-022-10198-9.

[13]. Desjardins, C. D., & Bulut, O. (2018). *Handbook of Educational Measurement and Psychometrics Using R*. CRC Press.

[14]. Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory. Principles and applications*. Springer Dordrecht.

[15]. Hulin, C.L., Drasgow, F., Parsons, C.K. (1983). *Item Response Theory: Application to Psychological Measurement. in Dorsey professional series.* Dow Jones-Irwin.

[16]. Kan, A & Bulut, O. (2014). Examining the Relationship Between Gender DIF and Language Complexity in Mathematics Assessments. *International Journal of Testing*, *14*(3), 245–264.
Doi: 10.1080/15305058.2013.877911

[17]. Zawistowska, A. (2017). Gender Differences in High-Stakes Maths Testing. Findings From Poland. *Studies in Logic, Grammar and Rhetoric*, *50*(1), 205–226. Doi: 10.1515/slgr-2017-0025.

[18]. Walker, C. M. & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*(2), 147–163.
Doi: 10.1111/j.1745-3984.2001.tb01120.x

[19]. Eren, B., Gündüz, T., Tan, S. (2023). Comparison of Methods Used in Detection of DIF in Cognitive Diagnostic Models with Traditional Methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 76–94.
Doi: 10.21031/epod.1218144