Analysis of Model fit and Item Parameter of Mathematics National Examination Using Item Response Theory

Aristiawan, Heri Retnawati, Edi Istiyono
Yogyakarta State University
aristiawan.2017@student.uny.ac.id

Abstract: This study aims to determine the compatibility of the item response theory models in Mathematics National Exam and to find out the estimated parameters of the items. Qualitative analysis was carried out on the responses of 3223 students in Yogyakarta. The analysis was carried out using statistical and graph methods. Based on the results of the model suitability analysis, it was found that the model suitable for the Mathematics National Examination was a three-parameter logistic model (3 PL). The difficulty in the range of -0.887 to 1.013, discrimination between 0.78 - 3.551 and pseudo guessing between 0.181 to 0.5.

Keywords: item response theory, model fit, item parameter, final examination

Programs to improve the quality of education must be designed based on accurate data because a program that is not based on the right data will not give optimal effect. This accurate data can be obtained through a good process. Mardapi (2017) states the quality of education can be improved by increasing the quality of learning and the assessment system. The results of proper system assessment will be feasible based on making further decisions.

In the education system in Indonesia, it is known as the National Examination which is a national assessment that is used as a standard measurement the achievement of graduate competence. Based on Government Regulation no 13 the Year 2015, national examinations are functioned as one of the considerations for mapping the quality of programs and/or educational units, the basis for selection to the next level of education, and fostering and providing assistance to education units in their efforts to improve education quality.

National Examination has a strategic function in the education system in Indonesia. Therefore, the implementation of this National Examination needs to be handled optimally from the planning to the reporting. One of the demands that must be fulfilled by the National Examination is that the results describe the abilities of students accurately. An assessment is called accurate if the results of the assessment contain errors as small as possible. To get a good result like that, a good quality test instrument is needed. The quality of a test instrument can be seen in its validity, reliability and item parameters.

There are two types of approaches for estimating item parameters, namely classical test theory and item response theory. Classical test theory is seen as having weaknesses. The most critical weakness according to Hambleton, Swaminathan, & Rogers (1991) is that the characteristics of examinees and test characteristics cannot be separated, each of which can be interpreted only in another context. The ability of the examinee is only determined by the test. When the test is difficult, the examinee will appear to have low ability, and when the test is easy, the examinee will seem to have higher ability. In other words, the estimated item parameters depend on the subject and vice versa. The item characteristics will change when the examinees change, and the characteristics of the test examinee will change when the characteristics of the items change. In this case, the classical test theory cannot be used as a standard because the results of the assessment depend on the subject.

The item response theory is a solution to overcome the weaknesses that exist in the classical test theory because the item responses theory has the concept of releasing the linkages between items and samples. The characteristics of the examinees will remain the same even though they work on items with diverse characteristics, and on the contrary, the characteristics of the items will remain the same even though they are done by examinees with different abilities. According to Hambleton et al., (1991) , item responses theory rests on two basic postulates: (a) the performance of examinees on test items can be predicted (or explained) by a series of factors called trait, latent nature, or ability; and (b) the relationship between the performance of the examinee items and the set of characteristics underlying the performance of items can be explained by functions that increase monotonically called the item characteristic function or item characteristic curve (ICC). This function explains that when the ability increases, the probability of the respondent answering correctly to an item increases. In Figure 1 it can be seen that groups of examinees with higher abilities will have a greater probability of answering correctly than groups of examinees with low abilities.

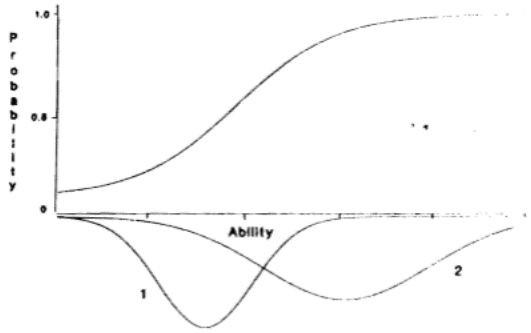

Figure 1. Item characteristics curve and ability distribution in two groups of examinees

The function of item response theory can be applied when the model used has a match with the test data (Hambleton et al., 1991). Stone & Zhang (2003) stated that the use of estimated item parameters could be disrupted when a model does not fit the data. Hambleton et al., (1991) describes several logistic models in item response theory, namely one-parameter logistic model (1PL), two-parameter logistic model (2PL), and three-parameter logistic model (3PL). Each model has a different number of item parameters. The parameters of this item serve as forming the item response function.

The one-parameter logistic model is an item response theory model which has only one parameter, i.e., difficulty (b). In this model, it is assumed that the ability of the examinees is only influenced by the difficulty of the item. The items can be said to be good if the difficulty is in the range - 2 which means it is easy to +2 which means difficult. The function of the 1 PL model can be seen in Equation 1

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

The two-parameters logistic model has an item parameter in the form of difficulty (b) and item discrimination (a) whose value is between 0 and 2. In the Item Characteristic Curve, discrimination is indicated by the slope of the curve. Items which have high discrimination will have a steep slope of the characteristic curve. Items that have steep slopes will be better able to distinguish high-skilled examinees with low-ability examinees. The function of the 2 PL model can be seen in Equation 2.

$$P_i(\theta) = \frac{e^{a(\theta - b_i)}}{1+e^{a(\theta - b_i)}} \quad (2)$$

The three-logistic parameters model has parameters of difficulty (b), discrimination (a) and pseudo guessing (c). The pseudo guessing parameter states the chances of participants with low ability to correctly answer a difficult item by guessing. The value of c ranges from 0 and 1. An item is said to be good if the value of parameter c is not more than 1 / k, with k being the number of choices. The functions of the 3 PL model can be seen in Equation 3.

$$P_i(\theta) = c + (1 - c)\frac{e^{a(\theta - b_i)}}{1+e^{a(\theta - b_i)}} \quad (3)$$

According to Retnawati (2014), two ways can be used to determine the suitability of the model, namely the statistical method and the graph method. The statistical method is done by calculating the chi-square value and then comparing it with the chi-square value of the table, or by reviewing the probability value (significance). The item is said to fit a model if the chi-square calculation does not exceed the chi-square value in the table or sig> α value. Meanwhile, the analysis using the graph method is done by looking at the distribution of data

on the item characteristic curve. Through this curve, it can be seen how precise the data distribution is compared to the model. The model is said to be suitable if the distance of the point with the match line is close (Retnawati, 2014).

This study aims to determine the suitability of the model from the Mathematics National Exam of junior high school in the academic year 2014/2015 and to find out the item parameters of the question based on the suitable model.

METHOD

This research is qualitative research with analysis document method. The document is the answers of junior high school students on the questions of packages 5 Mathematics National Exam in the academic year 2014/2015. Data were 3223 student's response and collected from 4 districts and 1 municipality in the province of the Special Region of Yogyakarta. Data is collected through the Puspendik document of National Education Ministry of Education. This research was conducted in the Special Region of Yogyakarta (DIY) from February to March 2019.

Qualitative analysis is carried out by carrying out factor analysis to test the unidimensionality assumptions using SPSS. The analysis was then continued with determining the suitability of the model using Bilog-MG 3.0. Model suitability analysis is done using the statistical method and with the graph method. After determining the suitability of the model, the analysis of the items is continued by looking at the results of the items parameter estimation based on the suitable model. The results of the item parameter analysis can be seen in the output Bilog MG 3.0 phase 2. The Threshold column shows the difficulty of the item, the Slope shows the discrimination and Asymptote states the pseudo guessing parameter.

RESULTS AND DISCUSSION

Results

First, the analysis was carried out by proving the assumption of item response theory, namely the unidimensionality and local independence. The assumption of unidimensionality in this study is proven through factor analysis using SPSS.

Factor analysis was carried out by first doing a feasibility test analysis, namely KMO-MSA test and Bartlett test. The results of the feasibility test can be seen in the Table 1 below.

Table 1. KMO dan Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.984
Bartlett's Test of Sphericity	Approx. Chi-Square	41263.883
	Df	780
	Sig.	.000

The results of factor analysis through SPSS can be seen in the eigenvalue section in Table 2 below.

Table 2. Eigenvalue

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	12.767	31.917	31.917
2	1.634	4.084	36.001
3	1.034	2.586	38.587

This eigenvalue can then be presented in the scree plot as follows:

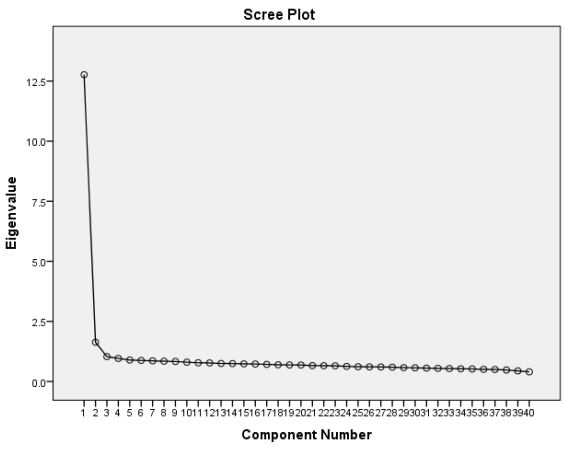

Figure 2. Scree plot Factor Analysis

The results of the suitability of the model with the chi-square comparison method can be presented in Table 3

Table 3. Model Fit Based on Comparison of Chi-Square Value

Model IRT	Total Items	Items
1 PL	2	31, 37
2 PL	6	4, 5, 12, 14, 16, 38
3 PL	21	2, 4, 5, 6, 8, 11, 12, 14, 15, 16, 17, 20, 21, 22, 25, 29, 31, 32, 36, 38, 40

The results of the model match with the graph method can be presented in Table 4.

Table 4. Model Fit Based on Graphic Method

Model IRT	Total Items	Items
1 PL	2	31, 37
2 PL	17	5, 7, 10, 13, 14, 16, 17, 18, 19, 23, 24, 26, 33, 34, 35, 38, 39
3 PL	21	1, 2, 3, 4, 6, 8, 9, 11, 12, 15, 20, 21, 22, 25, 27, 28, 29, 30, 32, 36, 40

Discussion

Unidimensionality

Before conducting a model compatibility analysis, an analysis is carried out first to prove the assumption of the item response theory, namely unidimensionality and local independence. Unidimensionality means that each item tests only measures one ability (Retnawati, 2014).

Factor analysis was carried out by first doing a feasibility test analysis, namely KMO-MSA test and Bartlett test. The KMO-MSA test aims to see the adequacy of the sample, while the Barlett's test serves to prove the homogeneity of the data. Factor analysis can be continued if the Kaiser-Meyer Olkin (KMO) -MSA value> 0.5 and significant Barlett's test <0.05 (Hair, Black, Babin, & Anderson, 2009). Based on Table 1 it can be seen that the KMO-MSA value is 0.984 and significant Bartlett test is 0.000. This means that the sample used has met the requirements for the adequacy of the sample and the data is homogeneous data, so that factor analysis can be done.

The results of factor analysis through SPSS can be seen in the eigenvalue section in Table 2. Eigenvalue whose value is more than 1 means 1 factor so that the question about the packages 5 Mathematics National Exam of junior high school in the academic year 2014/2015 has 3 factors. Of these three factors, there are 38,587% of variance that can be explained. This eigenvalue then can be presented in the scree plot at Figure 2.

Scree plots from the factor analysis showed a very sharp decrease between factor 1 and factor 2, the Eigenvalue then began to slope at the 3rd factor so that the scree plots almost formed a right angle. This indicates that there is only 1 dominant factor in the test set so that the assumed unidimensionality is fulfilled.

Local Independence

The assumption of local independence means that the test participant's response to one pair of pairs will be independent of conditions when the conditions of the aspects that influence them do not change.

According to DeMars (2010), if the assumed unidimensionality is met, the assumption of local independence is also fulfilled. So that all the assumption of the item response theory has been fulfilled and can be continued to analyze the suitability of the model.

With the fulfillment of the item response theory assumption, analysis of item compatibility can be continued.

Model Suitability with Statistics Methods

The statistical method is done by comparing the chi-square calculated with the value of the chi-square table on certain degrees of freedom. Item is fit with the model if the value of the chi-square count does not exceed the chi-square value of the table. By using the statistical method, it is found that the number of items that match the model 1 PL is 2 items, model 2 PL is 6 items, and model 3 PL is 21 items. (Table 3)

Model Suitability with Graphics Methods

As a comparison of the results of statistical method analysis, a suitability analysis of the model was carried out using the graph method. The graph method conducted by looking at the distribution of data on the item characteristic curve. The item is fit with the model if the distance of the point representing the data is close to the match line. For example, consider the item characteristic curve in number 9 which is analyzed by the 1 PL model, 2 PL and PL as follows

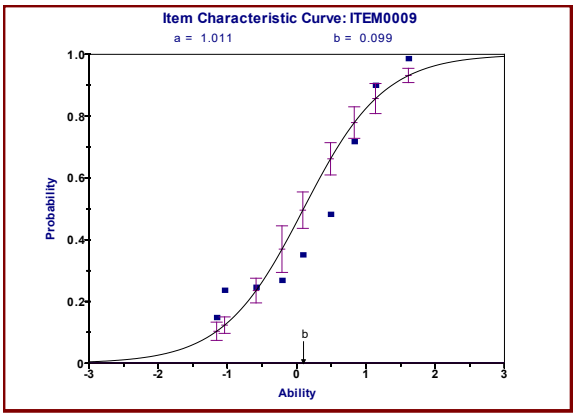

Figure 3. ICC of item number 9 with model 1 PL

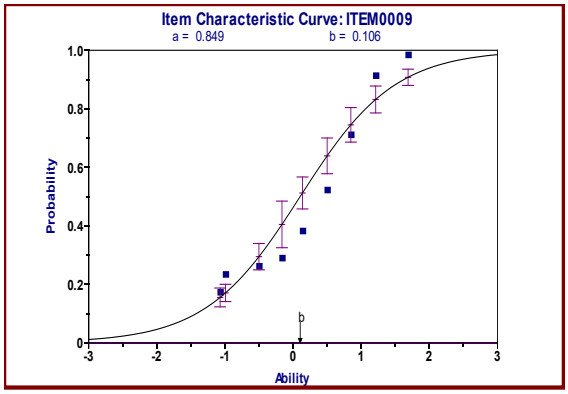

Figure 4. ICC of item number 9 with model 2 PL

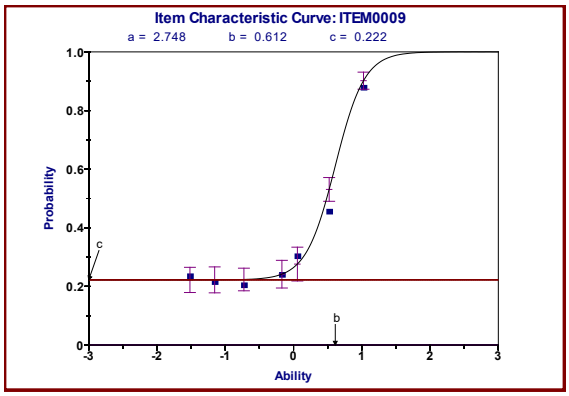

Figure 5. ICC of item number 9 with model 3 PL

Based on the three figures above, it can be seen that the distance of data distribution with the line match model number 9 will be more tightly if analyzed by 3 PL model compared with the 1 PL or 2 PL model, so that item number 9 is more appropriate if analyzed using the model 3 PL. Analysis of model compatibility with the graph method shows that there are 2 items that match the item response theory 1 PL model, 17 items that match the 2 PL model, and 21 items that match the 3 PL model.

Based on the suitability analysis of the model with the statistical method and the graph method, the same results were obtained, that the item response theory model that fits the packages 5 Mathematics National Exam of junior high school in the academic year 2014/2015 is model 3 PL while the model that produces the least suitable item is model 1 PL.

Mardapi (1998) states that 1 PL model is an item response theory model that has the most assumptions compared to 2 PL and 3 PL model. The 1 PL model will only estimate the difficulty of the item, while the item discrimination parameter must be assumed to be the same, and the pseudo guessing parameter must be assumed to be zero. The 2 PL model has an item parameter in the form of difficulty and discrimination, while the pseudo guessing parameter is assumed to be zero. Meanwhile, the 3 PL model has no assumptions regarding the item parameters, so that all large item parameters, both the difficulty, discrimination, and pseudo guessing need to be estimated. Because the 1 PL model has the most assumptions, this 1 PL model will produce a few suitable items.

The results of the analysis on the packages 5 Mathematics National Exam of junior high school in the academic year 2014/2015 showed that only a few items were suitable for the model. This is due to a large number of data responses used for analysis (more than 3,000 test participants). Retnawati (2014) states that the number of respondents will influence the value of chi-squared calculated. The greater the number of response of the test participants used for analysis, the greater the value

of the chi square calculated. This is also supported by the results of Fan, Thompson, & Wang (2009) that $\chi 2$ (chi squared) is a direct function of sample size, so the value of $\chi 2$ will be influenced by sample size. With the increasing value of the calculated chi-square, it will increase the chance to reject the item hypothesis thus minimizing the possibility of a model match.

Meijer (1996) mentions seven examinee behaviors when the test that causes the item does not match the data. The seven behaviors are a) Sleeping Behavior, an examinee has trouble in starting the task, and after having adapted, he does not check the answers; b) Guess behavior, where examinees with low abilities unexpectedly respond correctly to difficult items; c) cheating behavior; d) Plodding Behavior, namely examinees who have not finished working on the problem; e) Error Alignment, occurs in test examinees who are not careful in responding to the answer sheet; f) too creative, that is, examinees who interpret items in unusual or overly creative ways; g) Deficiency of Abilities, occurs when the problem measures two different abilities.

The item analysis was then continued to estimating the item parameter by referring to the 3 PL model, which means that the estimated parameters of the items were in the form of difficulty, discrimination, and pseudo guessing. From the analysis performed, it was found that the difficulty level of the problem was in the range of -0.887 to 1.013, the discrimination was in the range 0.78 to 3.551, and the pseudo guessing was in the range 0.181 to 0.5.

When referring to the item quality criteria according to Hambleton, the level of difficulty is in the range of -2 to +2, the discrimination is in the range of 0 to 2, and the pseudo guessing is not more than 0.25 (1 / k). in the category of not good. The results of items quality can be seen in Table 5 below.

Table 5. Items Quality

Quality	Total Items	Items
Good	8	1, 14, 18, 23, 29, 31, 39, 40
Poor Good	32	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, 24, 25, 26, 27, 28, 30, 32, 33, 34, 35, 36, 37, 38

CONCLUSION

The unidimensional assumptions and local independence on the packages 5 Mathematics National Exam of junior high school in the academic year 2014/2015 have been fulfilled. Based on the analysis with the statistical method and the graph method, the item response theory model that fits the packages 5 Mathematics National Exam of junior high school in the academic year 2014/2015 is model 3 PL. From the analysis, it was found that the difficulty level of the problem was in the range of -0.887 to 1.013, the discrimination was in the range 0.78 to 3.551, and the pseudo guessing was in the range 0.181 to 0.5.

REFERENCES

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

Fan, X., Thompson, B., & Wang, L. (2009). Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes. *Journal Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56–83.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). Multivariate Data Analysis (7th Edition). New Jersey: Prentice Hall.

Hambleton, R. K., Swaminathan, H., & Rogers, D. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park. California: Sage Publications, Inc.

Mardapi, D. (1998). Analisis Butir dengan Teori Tes Klasik dan Teori Respons Butir. *Jurnal Kependidikan*, 25–34.

Mardapi, D. (2017). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Yogyakarta: Parama Publishing.

Meijer, R. R. (1996). Person-Fit Research : An Introduction. *Applied Measurement in Education*, *9*(1), 3–8. https://doi.org/10.1207/s15324818ame0901

Retnawati, H. (2014). *Teori Respon Butir dan Penerapannya*. Yogyakarta: Nuha Medika.

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.