DOI: 10.13189/ujer.2020.080646

The Characteristics of High-Risk Tryout Test Items for Indonesian Elementary Schools Students

Retno Widyaningrum[1,2,*], Badrun Kartowagiran[1], Jailani[3]

[1]Department of Educational Research and Evaluation, Post Graduate Program, Yogyakarta State University, Indonesia
[2]Department of PGMI, Faculty of Tarbiyah and Education, IAIN Ponorogo, Indonesia
[3]Department of Mathematics Education, Faculty of Math and Science, Yogyakarta State University, Indonesia

Received April 6, 2020; Revised April 28, 2020; Accepted May 27, 2020

Abstract One of the measurement forms applied in the educational field to know students' ability is a high-risk tryout test. This research examined the quality of Mathematics tryout test questions used by Indonesian teachers to prepare for the national examination as a high-risk examination. This study used a quantitative approach and analyzed based on item response theory to determine the fitness of the model and item eligibility. The data used were the answers of elementary school students who took the 2016 national examination tryout of mathematics subject with the total number of 924 students from 20 elementary schools in Indonesia. The mathematics subject consisting of 40 questions, which are adjusted to the blueprint of elementary school national examination, was chosen. The results obtained indicate that the mathematics tryout test items meet the assumptions of uni-dimension, local independence, and parameter invariance items. The logistic parameter model used was the 2-PL model, with 31 out of 40 eligible items of questions. The novelty of this research is the use of item response theory to analyze the quality of an instrument. The aim of the use of item response theory is to avoid examinee sample dependent and item sample dependent. The educational stakeholders can use this research as a reflection in preparing high-risk examination.

Keywords Mathematics National Examination Tryout, Assumption of Item Response Theory, Model Fitness, Item Classification

1. Introduction

Assessing or measuring students' abilities is part of the education system. One of the purposes of evaluating/measuring students' ability is to use it as teachers' reflection material to improve teaching and learning processes and conduct a remedial program for students. It is also used to know the progress or learning outcomes of each student for reporting to parents, determining class progress, and determining students' graduation. Further, it is also used to treat students, in teaching and learning situations, appropriately to their abilities or characteristics, to know the background (psychology, physical and environmental) of students who experience learning difficulties. The results can be used as a basis for solving students' learning difficulties.

Measurement, in the educational field, means measuring the attributes or characteristics of individual students. In this case, it is not the students who are measured, but their characteristics or traits. This is according to explanation [1], which defines measurement as assigning numbers to individuals in a systematic way that reflects the characteristics of individuals. The instruments that can be used to obtain information about the characteristics of student abilities can be in the form of questions.

The questions made by the teachers need to be analyzed whether or not they measure the right competencies. This should be done repeatedly at the end of the learning process, which will eventually be measured at the end of the study period. The end of the study period refers to the national examination that has been carried out in Indonesia. A national exam is done to measure and determine the ability of students at the end of their studies. To prepare the national exam and to get maximum results, many efforts have been done by both schools and individuals, such as a tryout test. Many institutions host a tryout test, starting from learner courses institution and even by the school itself. However, there has not been any research that investigates the tryout test items whether or not the items are following national examination standards and are capable of measuring students' abilities.

It is essential to ensure the quality of the tryout test items to be able to measure the students' ability to prepare for the National Examination. Generally, the ones who make the tryout test items are the team of teachers in the local area. In the implementation of the national examination tryout for Mathematics subject in elementary schools of Indonesian suburbs, a multiple-choice test was used to measure students' abilities. The test items compiled by the Teacher Team of a particular school group are used for elementary schools, which are incorporated in that school group without analyzing the quality of the test items and analyzing the fitness of the instrument with the ability of students.

Therefore, to overcome weaknesses in the preparation of the national examination tryout test as applied to the Mathematics tryout test in the particular district, it is crucial to conduct an analysis of the quality of Mathematics tryout tests compiled by the teachers of a specific school group, since the results of the tryout can describe students' passing rate. The analysis of the quality of this tryout test item is focused on the analysis of item response theory because the analysis has a mathematical model, which means that the opportunity for the test takers to have correct answers depends on the ability of the students and the characteristics of items [2]. Based on the explanation above, a study is needed to test the assumptions of the item response theory in the developed test, to test the fitness of the item response theory model, and to analyze the items whether or not they are feasible to prepare students for national examination.

2. Materials and Methods

The research method used in this study was a quantitative approach with the item response theory. The data used were the answers of elementary school students who took the 2016 national examination tryout of mathematics subject with the total number of 924 students from 20 elementary schools in Indonesia. The mathematics subject consisting of 40 questions, which are adjusted to the blueprint of elementary school national examination, was chosen.

The raw data which was collected based on selected tryout samples were processed and analyzed to see the met assumptions of the item response theory consisting of uni-dimension, local independence, and parameter invariance, testing the fitness of the logistical parameter model of the item response theory, i.e., whether the questions are developed according to 1-PL, 2-PL or 3-PL models. From the logistical parameter model that is met, then the magnitude of the discriminating power parameters, the level of difficulty, and pseudo-guessing that are known were used to determine whether the items in the tryout questions are good or not.

The software used to process data was SPSS 20, BILOG MG 3.0, and Microsoft Excel 2013. The SPSS program was used to test local uni-dimensional assumptions and independence with factor analysis. BILOG MG 3.0 and Microsoft Excel 2013 were used to test the assumptions of parameter invariance as well as to test the fitness of the model on the tryout questions that were tested and to analyze the good and bad items.

3. Results and Discussion

Examine the characteristics of the mathematics national examination tryout items for elementary schools, on a broad scale, is done by testing the assumptions of the item response theory, i.e., uni-dimension, local independence, and parameter invariance. Uni-dimension is to get each test item only measures one ability; local independence is to find out the factors that influence the constant tryout results; the subject's response to any item pair will be statistically independent of one another; while parameter invariance examines the characteristics of the item which do not depend on the parameter distribution of the ability of test-takers and the parameters that characterize test takers do not depend on the characteristics of the items [2].

This article also analyzes the fitness test of the mathematics national examination tryout model for elementary schools that have been developed with 1-PL, 2-PL, and 3-PL models. From these results, parameters that fit the model, and the characteristics of both good and bad items from the developed instrument will be obtained. The following are the results of the discussion of the item response theory analysis that has been carried out.

3.1. The Assumptions of Item Response Theory Test

The quality of the questions will affect the accuracy to determine the ability of respondents. Detecting the quality of questions can be done through theoretical and empirical analysis. In general, this empirical item analysis can be divided into two, i.e., the classical test theory approach and Item Response Theory (IRT).

The classical test theory, or pure classical score theory, is based on an additive model, i.e., the observed score is the sum of the actual score and measurement error score [3]. Classical test theory has developed widely and has been the mainstream among psychologists and education experts, as well as other behavioral studies, for 20 decades [3]. Classical test theory has weaknesses because it is examinee sample dependent and sample dependent items [4]; [5]; [6]; [7]. Other weaknesses in classical test theory are that the parameters in classical test theory are the characteristics of items that depend on the sample group used, besides that, classical test theory also requires the equality of measurement error for all subjects involved in the test, parallel definitions in classical test theory is also practically very difficult to fulfill. The weaknesses of the classical test theory triggered a more adequate new theory, i.e., The Item Response Theory (IRT) or Latent Traits Theory (LTT).

This study used the IRT approach. According to [2], one of the IRT assumptions is that a uni-dimensional test is carried out using factor analysis, to see the Eigenvalues on the covariance-variance matrix of inter items. The data analysis with factor analysis was preceded by an analysis of sample adequacy using SPSS. The results of the analysis with SPSS are obtained as follows.

Table 1. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.827
Bartlett's Test of Sphericity	Approx. Chi-Square	17763.545
	df	780
	Sig.	0

The Chi-Square value indicates the analysis of the sample adequacy in the Bartlett test of 17763.545 with 780 degrees of freedom and a *p*-value of less than 0.01. This shows that the sample size of 924 in this study was sufficient to test the assumptions of the item response theory, namely uni-dimension. To find out the uni-dimensionality, SPSS is used with a scree plot to know the Eigenvalues as follows.

Eigenvalues are presented by the scree plot in Figure 1 above. Based on the results of the scree plot, it appears that the eigenvalue began to run off at the 13th factor. This shows that there is more than 1 dominant factor in the mathematics national examination tryout for elementary schools. Still, the other 11 factors do not contribute significantly to the variance component that can be explained. The conclusion is that the developed instrument measures have at least 2 factors with the first factor as the dominant factor.

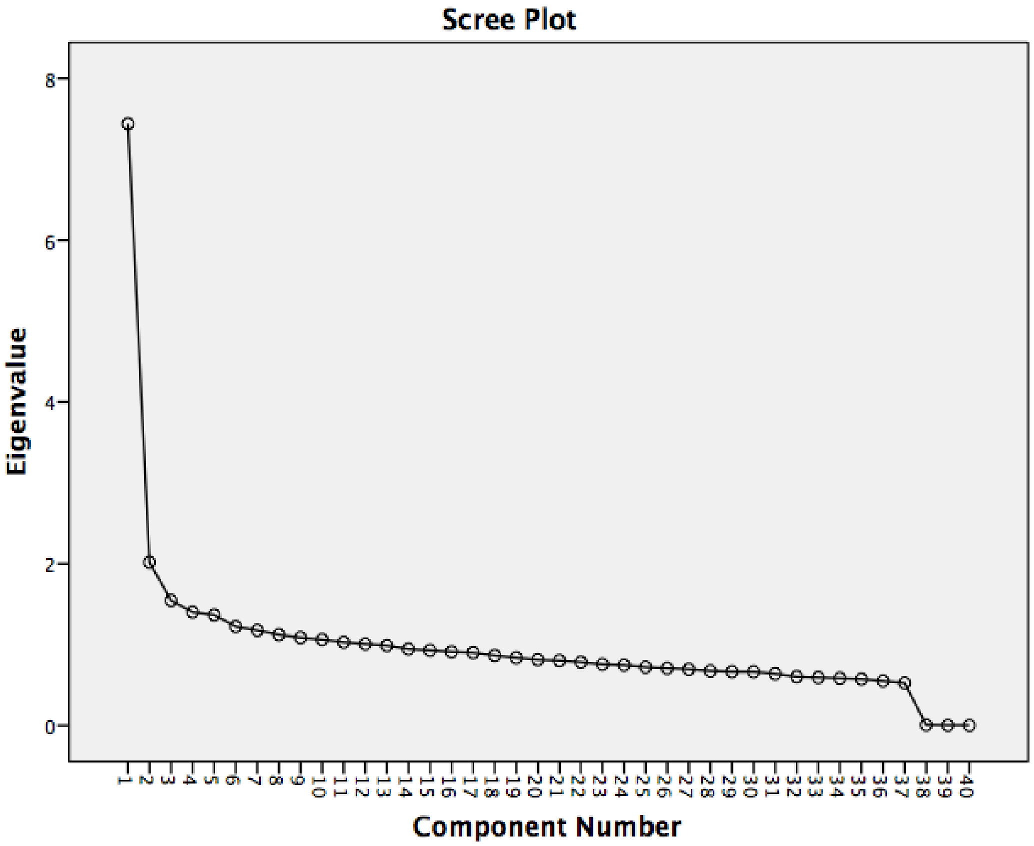

Figure 1. The Results of Scree Plot Factor Analysis

Another way that can be done is by looking at the percentage of variance, which is higher than 20% or a comparison of the first Eigenvalues with the second ones of 5 or 4 [8]. Based on data analysis with the SPSS 22 program, the following results were obtained.

Table 2. Eigenvalues and the Variance Components of the Results of Factor Analysis

Component	Initial Eigenvalues		
	Total	% of variance	Cumulative %
1	7.439	18.597	18.597
2	2.019	5.048	23.645
3	1.547	3.867	27.512
4	1.403	3.507	31.019
5	1.368	3.419	34.438
6	1.223	3.059	37.497
7	1.177	2.942	40.438
8	1.123	2.808	43.246
9	1.087	2.717	45.963
10	1.063	2.657	48.620
11	1.031	2.576	51.196
12	1.009	2.522	53.719
13	.989	2.473	56.191
14	.946	2.366	58.557
15	.930	2.325	60.883
16	.912	2.280	63.163
17	.901	2.253	65.416
18	.866	2.165	67.581
19	.838	2.095	69.676
20	.814	2.036	71.712
21	.803	2.007	73.719
22	.783	1.957	75.677
23	.756	1.891	77.567
24	.747	1.868	79.435
25	.722	1.805	81.240
26	.708	1.770	83.010
27	.696	1.739	84.748
28	.675	1.687	86.436
29	.666	1.664	88.100
30	.665	1.662	89.762
31	.641	1.602	91.364
32	.605	1.512	92.876
33	.595	1.486	94.362
34	.585	1.463	95.825
35	.575	1.438	97.263
36	.552	1.381	98.643
37	.527	1.318	99.961
38	.008	.021	99.982
39	.004	.010	99.993
40	.003	.007	100.000

From table 2 above, it is seen that there are two high Eigenvalues, 7.439, and 2.019. Those two values appear to be much higher compared to other Eigenvalues, but the dominant one is 7.439, so, the dominant factor is the first factor. It is also following the explanation of [9];[10] that if there is only one dominant factor, then an instrument fulfills the uni-dimensional nature.

The second assumption is local independence as evidenced by the uni-dimensionality of participant response data on the mathematics national examination tryout for elementary schools, after the uni-dimensional assumption test, there is one dominant factor which means that the instrument measures one dimension of the ability to do national mathematics examination. [11] states that the results of the uni-dimensional analysis indicate that there is no correlation between the participant's response to an item to another item, or it can be said that the assumption of local independence has been fulfilled.

The third assumption is the invariance of item parameters and capability parameters. This assumption of item parameter invariance is evidenced by estimating item parameters in different groups of test-takers based on gender. In this article, the assumptions are proven by estimating the parameters of the items, which include discriminating power (a), difficulty level (b), and pseudo-guessing (c) in the group of test participants depicted in scatter plot diagrams [11].

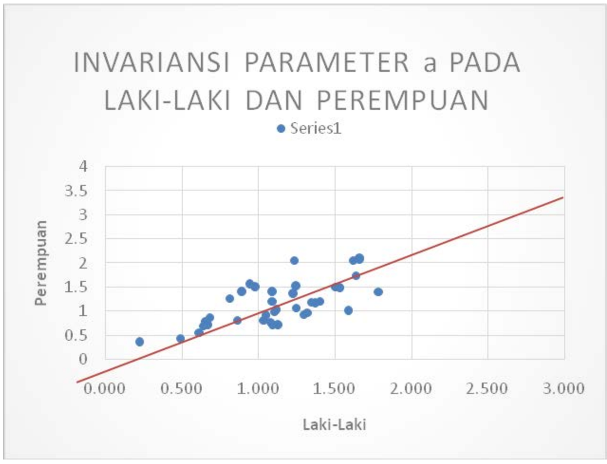

Figure 2. Parameter Invariance of Discriminating Power Based on Male and Female Groups

From the scatter plot above, it can be interpreted that the discriminating power parameters in the male and female groups of each point are relatively close to the slope 1. This shows that there is no variation in the estimated parameter results in the female and male groups.

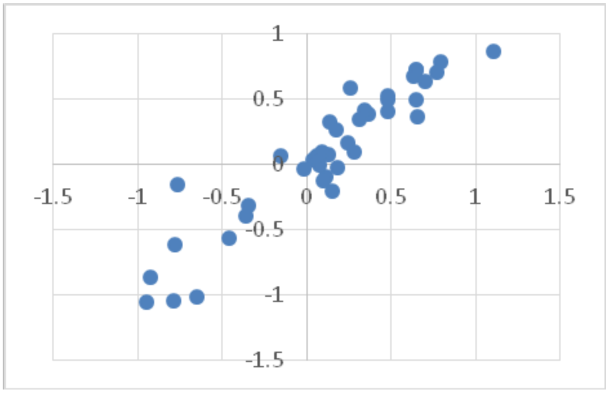

Figure 3. Parameter Invariance of Difficulty Level Based on Male and Female Groups

The scatter plot for the parameter (b), the level of difficulty estimated in the group of males and females, is depicted in figure 3. Based on the scatter plot, it was obtained that each point is relatively close to the line with slope 1. This means that there is no variation in the parameter of the level of difficulty of the estimation results in the male and female groups. In other words, parameter invariance is fulfilled.

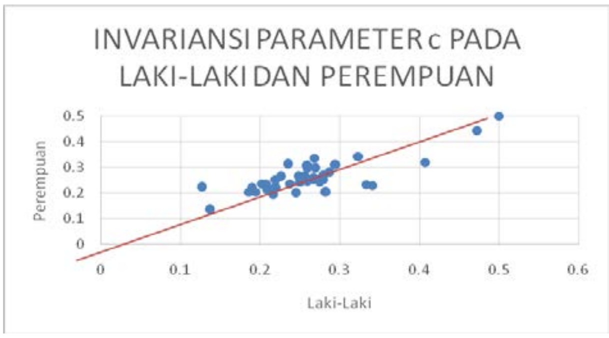

Figure 4. Parameter Invariance of Pseudo-Guessing Based on Male and Female Groups

The scatter plot for the parameter (c), pseudo-guessing estimated in the group of males and females, is depicted in Figure 4. Based on the scatter plot, it was obtained that each point is relatively close to the line with the slope of 1. This shows that there is no variation in the pseudo-guessing parameter estimation results in the male and female groups. In other words, parameter invariance is fulfilled.

The parameter invariance of ability is proven by estimating items parameters in different item groups based on odd and even numbers. In this article, the assumptions are proven by estimating item parameters that include discriminating power (a), difficulty level (b), and pseudo-guessing (c) in the Odd-Even group of clusters illustrated in scatter plot diagrams [11].

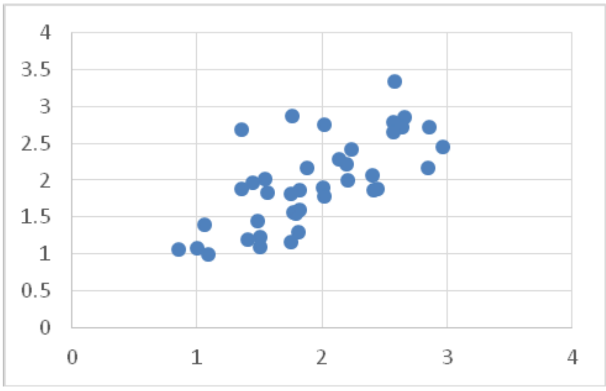

Figure 5. Parameter Invariance of Discriminating Power Ability Based on Odd-Even Groups

From the scatter plot diagram above, it was found that the mathematics national examination tryout items for elementary schools are spread out, but are not around the linear line, so the bias data is not invariant.

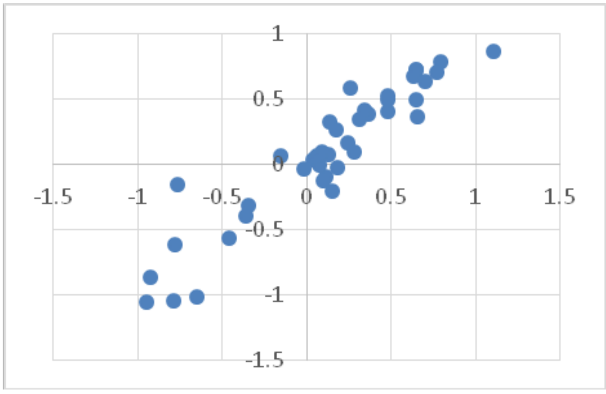

Figure 6. Parameter Invariance of Difficulty Level Ability Based on Odd-Even Groups

From the scatter plot diagram above, it can be evident that mathematics national examination tryout items for elementary schools are spread out, but are not around the linear line, so the bias data is not invariant with a linear line of 0.3.

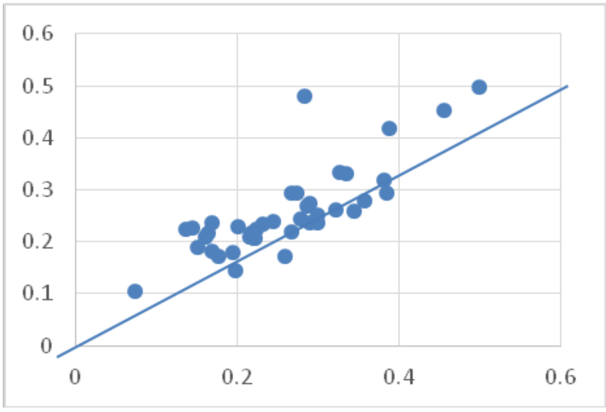

Figure 7. Parameter Invariance of Pseudo-Guessing Ability Based on Odd-Even Groups

From the scatter plot diagram above, it was found that the mathematics national examination tryout items for elementary schools are spread out, but are not around the linear line, so the bias data is not invariant.

Based on the test results above, the assumptions of item response theory in the tryout questions for elementary schools developed and tested on a broad scale have met all three characteristics, i.e., uni-dimension, local independence, and invariance of item parameters.

3.2. Fitness Test of Logistic Parameter Model

To test the fitness of the logistic parameter model, an analysis with item response theory with 1-PL, 2-PL, and 3-PL models can be used. The users of this theory need to choose whether the data analyzed is following one of the three models. Two ways can be used to determine the fitness of the analysis model that will be used, i.e., with the fitness of the model statistically and with the characteristic curve plot [11]. In this study, statistical model fitness was used.

From the statistical model selection, it was made the fitness of the three models based on the Chi-square value. The fitness of the models can be known by comparing the chi-square results of calculations with the chi-square table with a certain degree of freedom. The items fit with the model if the calculated chi-square value does not exceed the chi-square value of the table. The fitness is also known from the probability value (significance, sig.). If the sig. Value << X, then the item does not match the model [11]. In this study, the data obtained were analyzed using Bilog MG 3.0 software to obtain the Chi-Square value in determining the fitness of the model. The fitness of the PL 1 model is checked by testing at the probability/significance. If the sign value> a value = 0.05, then the model is fit or the model used is following the items developed, and vice versa, if the sign <value a = 0.05, the model used is not following the items developed.

Based on the analysis results, PL 1 model turns out to have 15 fit items, PL 2 model has 30 fit items, and 3 PL model has 25 fit items. From these results, most items are fit with Model 2 of Logistics Parameters. This shows that further item response theory analysis conducted in this study is to use the Model 2 Logistics Parameters.

3.3. Instrument Item Parameters and Item Classification of Good or Bad of 2 Logistics Parameters

The stages of the model fit test have been carried out as it turned out that the model's fit in the 2 Logistics Parameters was the most appropriate to the mathematics national examination tryout items for elementary schools that would be used to estimate the item parameters and abilities. The estimated parameters are the Discriminating Power (a), and Difficulty Level (b) with the help of Bilog MG 3.0 software.

From the analysis of items, 2 Logistics Parameters with Bilog-MG show that 7 items are not good because the level of difficulty of the items is less than 0.4, and the discriminating power is also low. Therefore, the seven items must be dropped/omitted. So, the questions for mathematics national examination tryouts are 31 good items and 9 bad items. The use of good items is recommended rather than bad, so, students can develop their thinking skills [9].

4. Conclusions

The assumptions of item response theory have been fulfilled by the tryout questions of Mathematics for elementary schools which have been tested on a broad scale consisting of uni-dimensions, local independence and invariance of item parameters. Uni-dimensional assumptions are indicated by factor analysis using SPSS and 1 dominant factor which is obtained from two factors resulting from the analysis. The assumption of local independence in the tryout questions of Mathematics is also fulfilled as a result of the previous assumption test, which states that the Mathematics tryout item is uni-dimensional. Meanwhile, item parameter invariance assumptions were also fulfilled based on the results of the analysis of student responses in different gender groups.

Based on the results of the model fit test with statistical analysis of significance values to determine which model is following the 2016 Mathematics tryout test, the 2 logistic parameter model is the fittest model with the magnitude of the discriminating power parameters (a) is -5.470 to 2.179 and the difficulty level parameter (b) is -1.087 to 1.926. From the known item parameters, it is obtained the results of the classification of good and bad items with 31 eligible items and 9 improper items.

REFERENCES

[1] M. J. Allen and W. M. Yen, *Interoduction to measuring*. Monterey, CA, 1979.

[2] R. K. Hambleton and H. Swaminathan, *Item Response Theory*. Boston, MA: Kluwer Inc., 1985.

[3] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologist*. NJ: Lawrence Erlbaum Associates Inc, 2000.

[4] X. Fan, "Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Response Person Statistics," *Educ. Psychol. Meas.*, vol. 58, no. 3, pp. 357–381, 1998.

[5] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamental of item re-sponse theory*. Newbury Park, CA: Sage Publication Inc., 1991.

[6] R. K. Hambleton, F. Robin, and D. Xing, "Item Response Models for the Analysis of Educational and Psychological Test Data," in *H. E. Tinsley, & S. D. Brown, Handbook of applied multivariate statistics and mathematical modeling*, San Diego, CA: Academic Press, 2000, pp. 553–581.

[7] F. M. Lord, *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers, 1980.

[8] C. S. Wells and U. Purwono, "Item response theory: polytomous respons IRT models and application," *Conferance of education*, Yogyakarta, 2008.

[9] M. Zainudin and E. Istiyono, "Scientific Approach to Promote Response Fluency Viewed from Social," *Eur. J. Educ. Res.*, vol. 8, no. 3, pp. 801–808, 2019.

[10] M. Zainudin, B. Subali, and Jailani, "Construct Validity of Mathematical Creativity Instrument : First-Order and Second-Order Confirmatory Factor Analysis," *Int. J. Instr.*, vol. 12, no. 3, 2019.

[11] H. Retnawati, *Validiti, reliability, and characterictic of item*. Yogyakarta: Parama Publising, 2016.