Journal of Mathematics Education
~JME~
Website: http://usnsj.com/index.php/JME
Email: editor.jme@usnsj.com

Development of Numbers Material Test Using the Item Response Theory (IRT) Approach for SD Students

AUTHORS INFO

Ahmad Rustam
Universitas Sulawesi Tenggara
Indonesia
ahmad.rustam1988@gmail.com

Kasmawati
Universitas Sulawesi Tenggara
Indonesia
kasmawatidullah268@gmail.com

ARTICLE INFO

o-ISSN: 2528-2026
p-ISSN: 2528-2468
Vol. 6, No. 2, December 2021
URL: http://doi.org/10.31327/jme.v6i2.1585

Suggetion for the Citation and Bibliography
Citation in Text:
Rustam & Kasmawati (2021)

Bibliography:
Rustam, A., & Kasmawati. (2021). Development of Numbers Material Test Using the Item Response Theory (IRT) Approach for Elementary School Students. *Journal of Mathematics Education,* 6(2), 78-86. http://doi.org/10.31327/jme.v6i2.1585

Abstract

The purpose of the research is to produce a product in the form of a valid and reliable measuring instrument for student numeracy that can be used in schools and in the general public. The research stages will be carried out based on the test development design, namely Preparing Test Specifications, Preparing Test Items, Testing Test Items in the Field, Revision of Test Items, and Test Development. The question grid is based on the 2013 curriculum syllabus. The test was conducted on elementary school students. The response of the test results in the form of dichotomous data and analyzed using the item response theory (IRT) model with two logistical parameters (2PL), namely the level of item difficulty and item discriminating power. Estimation of item parameters and capability parameters using the BILOG MG program. Before doing item analysis with IRT. The results of the study contained 18 items that could be used to measure students' numeracy skills. Among these items are numbered questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18. meet the criteria of a good item including having a good difficulty level, then the distinguishing power of the item functions well and has good validity and reliability.

Keywords: Test development, Numbers material, IRT, Two logistic parameters

A. Introduction

The educational process cannot be separated from the evaluation process. As it is known that the evaluation process is the most important thing in a curriculum. The evaluation process is a way to know that the success or failure of a learning process. The evaluation process is an effort made to control the quality of education nationally. The importance of the evaluation process requires that related elements in education must try to carry out and improve the evaluation process. The evaluation process can be carried out properly and correctly by using the right evaluation tools, both in schools and in certain areas specifically for each particular learning achievement.

The number of test kits that can be used makes it easier for students to evaluate student learning achievements. However, the type of test that is practical and most often used is the multiple choice test. This is in line with the statement that the most frequently used or commonly used type of test is the multiple choice test (Kean & Reilly, 2014; Raykov et al., 2019). It is important to develop test tools to provide an overview or results that are truly able to provide a valid description of student learning outcomes according to their abilities. For this reason, it requires educators to be able and have reliable test equipment, in this case the test used has been tested and has good validity and reliability values, so that the test is feasible to use to provide an overview of students' actual abilities.

The development of tests that are often used in schools still uses classical theoretical concepts. In fact, there are still many teachers who use questions from textbooks that are already available, whose validity and reliability quality of the test questions are not yet known. In addition to these problems, the use of classical theory has several weaknesses, so there is a bias in determining students' abilities. There are two weaknesses of classical theory; a) measurement results depend on the characteristics of the test used, and b) item parameters such as discriminating power and level of difficulty depend on the ability of test takers (Embretson & Reise, 2013). In addition, measurement errors in classical theory that can be searched for are groups not individuals. Along with the development of science, the use of classical theory began to be abandoned and shifted to a modern theory known as item response theory or item response theory (IRT). The basic concept of this theory is that there are main assumptions, namely, a) local independence, where the opportunity to correctly answer one item is mutually independent, b) unidimensional, where the substance being measured is one dimension (Embretson & Reise, 2013; Sarea & Ruslan, 2019).

Thus, the purpose of this study is to produce a product in the form of a valid and reliable measuring instrument for student numbers that can be used in junior high schools (SD) and in the general public.

Item Response Theory (IRT)

Item response theory or as other names (IRT) is a modern theoretical model that uses the concept of probability. The basic idea of IRT lies in two postulates, namely, a) the performance of a test taker can be predicted by one trait, namely the latent trait, b) the relationship between the performance of the test item and the set of characteristics underlying the item's performance can be explained by the item characteristic function or item characteristic curve. (ICC) (Kean & Reilly, 2014; Raykov et al., 2019).

There are assumptions in the IRT concept that must be met in order to be used in analyzing both item parameters and capability parameters, namely unidimensionality and local independence. Unidimensional, namely a collection of questions in a test device must measure only one dimension (single dimension). If people's ability to a set of questions depends on their position in two or more non-identical dimensions, it is not possible to represent people's interactions with questions with one single parameter, namely ability (Embretson & Reise, 2013). Meanwhile, local independence, namely the opportunity to correctly answer one item with another is independent (Embretson & Reise, 2013; Sarea & Ruslan, 2019).

In response theory, the ability or ability item can be in the form of cognitive ability, language, or even non-cognitive abilities such as skills, honesty and others. To determine the test taker's ability in item response theory using two logical parameters, it can be modeled as follows,

$$P(y_{ij} \mid \theta_j) = \frac{1}{1+e^{Da_i[\theta_j+b_i]}} \quad \text{..... (1)}$$

dimana,

P(y|θ) : probability of answering correctly on the i-th question by participants with ability (θ)

D : 1,7

e : 2,718

a_i : power parameter difference i-th question (*slope*)

b_i : the parameter of the difficulty level of the i-th question (*threshold*)

c_i : the parameter of guessing the i-th question

Based on the formula above, the symbol P(y|θ) represents the probability of the test taker answering correctly, while the symbols , ai and bi, represent the ability parameter, the different power parameter and the difficulty level parameter.

Items with a small value of discrepancy parameter (a) are able to provide some information about the ability in a wide tetha (θ) range. On the other hand, items with a large value of the discrepancy parameter (a) provide strong information about the value of in the area near the value of the difficulty level parameter (b), but only provide little information about for areas that are far from the value of the difficulty level parameter (b) (Von Davier et al., 2019).

The second parameter used in IRT is the level of difficulty parameter (b). The difficulty level parameter (b) is the point on the ability scale where a test taker has the opportunity to answer correctly the item is 0.5 for the 1PL and 2PL models, while the 3PL model is $\frac{(1+c)}{2}$ (Hambleton & Jones, 1993; Sarea & Ruslan, 2019; Subali et al., 2019).

The third parameter in the IRT used in the 3PL model is the guess parameter (c) which is called the pseudo-level-chance parameter. The guess parameter (c) is interpreted as the probability of the correct answer from very low ability test participants (Subali et al., 2019). Other quantities in the IRT are item characteristic curve or other terms item characteristic curve (ICC) and test information. The item characteristic curve (ICC) is a description of the relationship between the performance of the subject on an item and the underlying latent device. In addition, the item characteristic curve can be explained as the relationship between the probability of answering correctly $P_i(\theta)$ with the participant's level of ability (θ). This relationship is a nonlinear relationship between the item scores on the ability (θ) as measured by the test.

According to Hambleton, the test information function is the sum of the item information functions of each of those that compose the test (Embretson & Reise, 2013).

The formulation of the information function of a test is stated by Hambleton, Swaminathan, and Rogers (1991) as follows: (Subali et al., 2019).

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \quad \text{.......} (2)$$

$I_i(\theta)$ is the item information of i. The amount of information from the test is the sum of the information for each item described in equation (2). According to Birbaun (Hambleton & Jones, 1993) that maximum information will be achieved when ci = 0 in other words there is no guess and max = bi. The test information function is useful if the test items match the model (Kean & Reilly, 2014; Subali et al., 2019).

B. Methodology

This research was carried out for approximately eight months starting from April 2019 to November 2019. The research location was at the Kolaka Regency Elementary School. This research method is a survey research, where data collection responses from elementary school students. The response data is the results of the Numbers material test which is then analyzed to develop a Numbers material test instrument that can later be used by elementary students and those who need the test.

The stages of the research were carried out based on the stages of test development as follows (Hambleton & Jones, 1993):

The first stage, Preparation of Test Specifications

At this stage, it begins with the identification process of Numbers material based on competency standards, basic competencies, and indicators of Numbers material based on the 2013 curriculum.

The second stage, Preparation of the Test Item Pool

For this stage the researcher analyzed various references to develop test items. Based on the existing material indicators, items are developed for each indicator.

The third stage, Testing the Test Items in the Field (Field Testing the Items)

This step is carried out after the items have been compiled in the form of a test package. This is done to find out whether the test instructions can be understood well and the item questions that do not have ambiguous instructions. This stage is carried out in small groups, namely one class consisting of ± 30 students.

The fourth stage, Revision of the Test Items

At this stage the items that have been responded to by students are analyzed based on the pattern of student responses by reviewing the question sentences, answer keys and distracting items. This analysis uses item response theory by looking at the results of the test item analysis on the value of item discriminating power (a) and item difficulty level (b).

The fifth stage, Test Development

In this process or step, data is collected in the field using a large sample. The test will be conducted on elementary school students in Kolaka district by taking test participants of ± 300 students. After the response data is obtained, it is analyzed using item response theory or often called IRT with two logistical parameters (2PL) with the help of BILOGMG software. The results of this study in the form of the results of the analysis of the test items, namely the discriminating power of items (a) and the level of difficulty of items (b), and the results of this analysis will provide data on the ability of Numbers for each student.

C. Findings and Discussion

This research has been carried out with the aim of obtaining a good test instrument and in accordance with the criteria of a valid and reliable instrument. For this reason, the following steps or process of item development are presented following the development design of Hambleton and Jones.

Preparation of Test Specifications

At this stage, it begins with the process of identifying Numbers material in grade V Elementary School (SD) based on competency standards, basic competencies, and indicators of Numbers material based on the 2013 curriculum. There are several basic competencies (KD) and materials that will be used as a proposal to determine indicators in making questions based on the applicable curriculum. The following is presented in table 1 of the material along with the question indicators:

Table 1. Basic Competencies and Materials for Numbers

No.	Material	Question Indicator
1	Number to the power of two	Recognize the meaning of the power of two of a number.
2	Finding a number to the power of two takes the square root.	Perform addition, subtraction, multiplication and division operations with double-digit numbers
3	Add two fractions with different denominators.	Explain the addition of two fractions with different denominators
4	Subtracting two fractions with different denominators	Explain the subtraction of two fractions with different denominators.
5	Multiplication of Fractions by Fractions	Explain and perform multiplication of fractions by fractions.
6	Multiply a fraction by a decimal	Determine the result of multiplying a fraction by a decimal.
7	Division of fractions by decimal	Do the division between ordinary fractions and decimal fractions.
8	Problems related to multiplication and division of fractions.	Solve problems involving multiplication and division of fractions.
9	Use of scale in plans and problems.	Use of scale in plans and problems.

Based on the results of curriculum analysis in grade V elementary school (SD) there are several materials that examine the numeracy abilities of Grade V elementary school children, including in general the material for numbers to powers, fractions, decimal numbers, scales and comparisons.

From the grid that has been developed with the subject teacher and then used as a basis for developing number ability test items.

Preparation of the Test Item Pool

For this stage the researcher analyzed various references to develop test items. Based on the material indicators that already exist on the grid that has been developed, the items for each indicator are developed.

Some considerations are made before developing the problem, especially certain materials. This has also been explained by Muhsetyo, et al (2007) some problems or difficulties that may be faced or experienced by students, namely difficulty using fractions or rational numbers to show comparisons in certain situations, difficulty expressing comparisons in the form of division and fractions, difficulty understanding relationships congruence in geometry with corresponding fractions to express comparisons, and difficulty understanding ascending and descending scales. In addition, another consideration put forward by Divine, Yandari, and Pamungkas that the fractional number meter is a learning material that contains abstract concepts, every new abstract concept that students learn needs to be given repeatedly and periodically. Of course, consideration of the results of previous research can be used as a basis for thinking in the development of questions.

Making number material questions is developed by making a question card design. So each question consists of one question card, where each question card contains the identities of the questions. With the meaning that each question card explains things including subjects, classes, semesters, curriculum types, basic competencies, materials, question indicators, cognitive levels or cognitive dimensions, and then descriptions of questions, answer keys, and the criteria for the results of the study of questions.

Field Testing the Items

This step is carried out after the items have been compiled in the form of a test package. This is done to find out whether the test instructions can be understood well and the item questions that do not have ambiguous instructions. This stage is carried out in small groups, namely one class consisting of ± 30 students. The students who were the respondents were students from SD Negeri 1 Laloeha. Students who respond are students who have studied and understood the material concept of numbers.

The item parameters consist of three types, namely the distinguishing power parameter with the notation "a", the difficulty level of the item with the notation "b", and guesses with the notation "c". The 3-parameter logistic model (L3P) consists of three parameters, namely discrepancy (a), level of difficulty (b), and guesswork (c). The model (L2P) consists of 2 parameters, namely difference power (a), difficulty level (b), and the guess parameter is considered to be zero (c=0). Meanwhile, the model (L1P) consists of 1 parameter, namely the level of difficulty (b), the different power parameter is considered to be constant equal to 1 (a=1), and the guess parameter is considered to be zero (c=0).

1) Item Difficulty Level Parameter or threshold (b)

The item difficulty level (bi) aims to obtain information about the suitability of the item with the model. The item difficulty level is a function of the item that describes a person's ability.

Theoretically, the parameter of item difficulty level moves from -∞ bi , in item response theory. However, in practice the item difficulty level ranges from -2 bi +2. An item with a difficulty level below -2 means that the item is in the low category. Meanwhile, an item with a level of difficulty exceeding +2 means that the item is in a difficult category.

The results of the item analysis based on the output of BILOG MG at the output of PH.2, information was obtained that the item difficulty level moved from 3,147 to 15,156. The classification of item difficulty parameters is presented in table 2 as follows:

Table 2. Difficulty of small group test items

No.	Item Difficulty Index (b_i)	Category	Total	Item Number
1	$b_i > +2$	Hard	18	1, ,2,3,4,5,6,7,8,9,10,11,12,13,14,15 ,16,17,18
2	$-2 \leq b_i \leq +2$	Currently	0	-
3	$b_i < -2$	Easy		-
4	Missing item		2	19,20
Total Items			20	

Based on table 2, information is obtained that about 95% of all items have a good level of difficulty. It can be concluded that about 95% of all test items are able to describe the function of students' abilities. Meanwhile, and 5% of the test items have a level of difficulty including the category of bad items.

2) Parameters of Grain Dissimilarity or slope (a)

The grain discriminating power parameter (ai) is the slope of the grain at the point of difficulty for each item on a certain ability scale. The greater the slope of the curve, the greater the value of the power difference.

Theoretically the value of the difference power moves from -∞ ai . However, in practice the value of the power difference ranges from 0 ai 2. The results of the analysis of the power difference using the BILOG MG program are displayed on the output of PH2. Based on the results of the analysis of the PH2 output on the slope column, it was found that the power difference was in the range of 0.175. The classification of grain discriminating power parameters is presented in table 3 as follows:

Table 3. Differential power of small group test items

No.	Parameters of Grain Dissimilarity or slope (a_i)	Kategori	Total	Item Number
1	$a_i > 2$	Not good	2	19, 20
2	$0 \leq a_i \leq 2$	Good	18	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18
Total Items			20	

Based on table 3, information is obtained that about 95% of all items have good discriminating power. It can be concluded that about 95% of all test items are able to describe the function of students' abilities. Meanwhile, 5% of the test items have a different power value that is not good.

Revision of the Test Items

At this stage the items that have been responded to by students are analyzed based on the pattern of student responses by reviewing the question sentences, answer keys and distracting items. This analysis uses item response theory by looking at the results of the test item analysis on the value of item discriminating power (a) and item difficulty level (b). Based on the results of the analysis obtained information that about 95% or 18 items of all items have a good level of difficulty. It can be concluded that about 95% of all test items are able to describe the function of students' abilities. Meanwhile, and 5% or 2 test items, namely numbers 19 and 20, and both test items have a level of difficulty including the category of bad items.

The test items that do not meet the criteria for model fit and the item difficulty level, namely numbers 19 and 20, come from the indicators of the use of the scale on the plan and the problems. This question indicator indicates that the teacher has not taught the material for the item, so that almost all students answered, but the results were not correct.

Judging from the facts in the field that the item is not worth testing, the researcher still tries to include it on a large-scale test, assuming that with a large number of students, they are able to provide a variety of responses.

After that, it was obtained that information that about 95% of all items had good discriminating power. It can be concluded that about 95% of all test items are able to describe the function of students' abilities. Meanwhile, 5% of the test items have a different power value that is not good.

Test Development

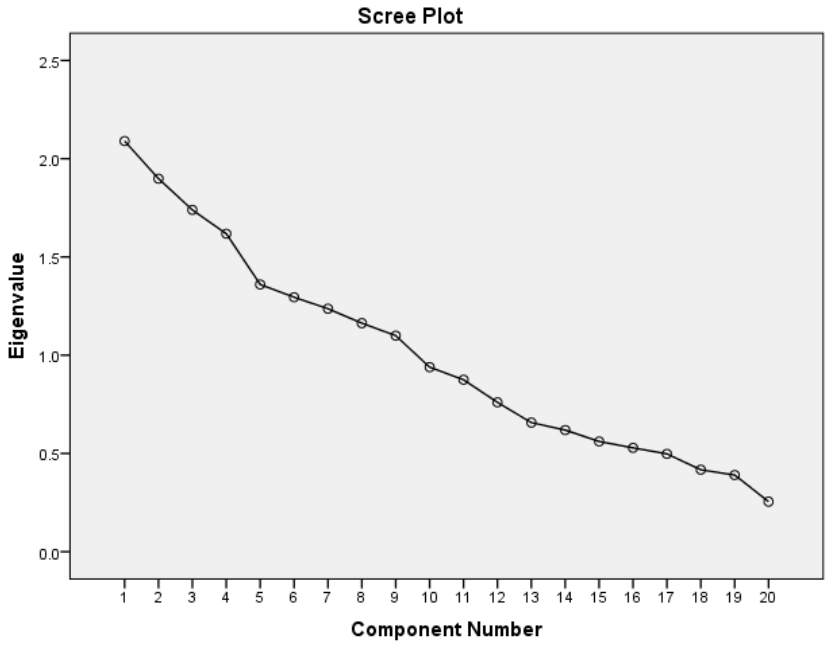
In this process or step, data is collected in the field using a large sample. The test was conducted on elementary school students in Kolaka district by taking test participants of ± 78 students. After the response data is obtained, it is analyzed using item response theory or often called IRT with two logistical parameters (2PL) with the help of BILOG MG software. The results of this study in the form of the results of the analysis of the test items, namely the

discriminatory power of items (a) and the level of difficulty of items (b), and these results will provide data on the ability of Numbers for each student. The following are the results of the analysis and discussion of the test item analysis.

1. Test the Assumptions of Item Response Theory
   a. Unidimensional test
      Based on the results of the item analysis using the factor test, the chi-square value in the Barlet test is 210.565 with 190 degrees of freedom and a p-value of 0.146, while the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) value is 0.442 which is greater than 0.05. . The following is presented in full in the image below:

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.442
Bartlett's Test of Sphericity	Approx. Chi-Square	210.565
	df	190
	Sig.	.146

      Apart from the KMO criteria above, it can be seen from the eigenvalues of the factors that the formed factors have eigenvalues that are not much different. Where the eigenvalues of each factor range from 2 and 1. This indicates that there is a dominant measurement result in measuring one particular dimension, so it can be concluded that this test instrument measures one dimension or unidimensionality. This is in line with what Hambleton and Swaminathan (1986) stated that the presence of a dominant dimension means that the test kit is unidimensional. Here's the full picture,



   b. Model Fit Test
      The results of the model fit analysis for items are very important to know to ensure that these items are feasible or not to be continued for further analysis. After analyzing the model test using BILOGMG, it was found that the significance value of the model fit on the PH2 output.
      Based on the results of the analysis in the image above, the significance value of the Threshold column (hard item level) there are 18 items whose significance value is greater than the value of = 0.05, thus from the 20 item numbers there are 18 items that match the model.

2. Analysis of Item Parameter Estimation Results and Ability Parameters

There are two parameters in the item response theory (IRT) estimation, namely the item parameter and the respondent's ability parameter. The respondent's ability parameter called theta "θ" states that a test taker is a trait with ability. Meanwhile, the item parameter states a

characteristic of the item through a mathematical model, namely a logistic model that fits the model.

The item parameters consist of three types, namely the distinguishing power parameter with the notation "a", the difficulty level of the item with the notation "b", and guesses with the notation "c". The 3-parameter logistic model (L3P) consists of three parameters, namely discrepancy (a), level of difficulty (b), and guesswork (c). The model (L2P) consists of 2 parameters, namely difference power (a), difficulty level (b), and the guess parameter is considered to be zero (c=0). Meanwhile, the model (L1P) consists of 1 parameter, namely the level of difficulty (b), the different power parameter is considered to be constant equal to 1 (a=1), and the guess parameter is considered to be zero (c=0). The analysis results from BILOG MG are displayed on the output PH.2.

1) Item Difficulty Level Parameter (b)

The item difficulty level (bi) aims to obtain information about the suitability of the item with the model. The item difficulty level is a function of the item that describes a person's ability.

Theoretically, the parameter of item difficulty level moves from -∞ bi , in item response theory. However, in practice the item difficulty level ranges from -2 bi +2. An item with a difficulty level below -2 means that the item is in the low category. Meanwhile, an item with a level of difficulty exceeding +2 means that the item is in a difficult category.

The results of the item analysis based on the output of BILOG MG at the output of PH.2, obtained information that the item difficulty level moved from 1.228 to 9.130. The classification of item difficulty parameters is presented as follows:

Table 4. Difficulty level of large group test items

No.	Parameters of Grain Dissimilarity or slope (a_i)	Category	Total	Item Number
1	$b_i > +2$	Hard	17	1,2,3,4,5,6,7,8,9,10,11,12,13,4,16,17,18
2	$-2 \leq b_i \leq +2$	Currently (Good)	1	15
3	$b_i < -2$	Easy	0	-
4	Missing Items		2	19,20
Total Items			40	

Based on table 4, information is obtained that about 95% of all items have a good level of difficulty. It can be concluded that about 95% of all test items are able to describe the function of students' abilities. Meanwhile, 5% of the test items have a level of difficulty including the category of bad items.

2) Parameter of Grain Difference (a)

The grain discriminating power parameter (ai) is the slope of the grain at the point of difficulty for each item on a certain ability scale. The greater the slope of the curve, the greater the value of the power difference.

Theoretically the value of the difference power moves from -∞ ai . However, in practice the value of the power difference ranges from 0 ai 2. The results of the analysis of the power difference using the BILOG MG program are displayed on the output of PH2. Based on the results of the analysis of the PH2 output on the slope column, it was found that the difference power was in the range of 0.257. The classification of the grain discriminating power parameters is presented as follows:

Table 5. Difficulty level of large group test items

No.	Parameters of Grain Dissimilarity or slope (a_i)	Kategori	Total	Item Number
1	$a_i > 2$	Not good	2	19, 20
2	$0 \leq a_i \leq 2$	Good	18	1,2,3,4,5,6,7,8,9,10,11,12,13,14, 15,16,17,18
Total Items			20	

Based on table 5, information is obtained that about 95% of all items have good discriminating power. It can be concluded that about 95% of all test items are able to describe the function of students' abilities. Meanwhile, 5% of the test items have a different power value that is not good.

3) Test Participants Ability Parameter (θ)

The ability parameter (θ) describes the characteristics of the test taker's ability. The estimation of the ability of the test takers can be seen in the PH3 output of the BILOG MG program analysis.

Based on the output of PH3, the average value of students' abilities is -0.0001 and the empirical reliability value is 0.1666. Taking into account the average value of students' abilities which are minus, it can be concluded that most students tend to have low abilities (θ)..

D. Conclusion

The purpose of this study was to obtain a standard instrument for elementary students' number material. The conclusion of this study is that of the 20 questions developed to measure elementary students' number material, there are 18 questions that can be used to measure students' numeracy skills. Among these items are numbered questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18. meet the criteria of a good item including having a good difficulty level, then the distinguishing power of the item functions well and has good validity and reliability.

E. References

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Hambleton, R. K., & Jones, R. W. (1993). *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*. Educ. Meas.

Ilahiyah, N., Yandari, I. A. V., & Pamungkas, A. N. (2019). Pengembangan Modul Matematika Berbasis Pakem Pada Materi Bilangan Pecahan Di SD. *Jurnal Pendidikan dan Pembelajaran Dasar, 6*(1), 49-63.

Irvine, S. H., & Kyllonen, P. C. (2013) *Item generation for test development* . London: Routledge.

Kean, J., & Reilly, J. (2014). *Item response theory, in Handbook for clinical research: Design, statistics and implementation*.

Maddalora, A. L. M. (2019). Personalized learning model using item response theory," *Int. J. Recent Technol. Eng, 8*(1) Special Issue 4, 811–818.

Muhsetyo, G., dkk. 2007. *Pembelajaran Matematika SD*. Jakarta: Universitas Terbuka.

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., & M, H. (2019). On true score evaluation using item response theory modeling. *Educ. Psychol. Meas*, *79*(4), 796–807.

Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory vs Item Response Theory? *Didakt. J. Kependidikan*, *13*(1), 1–16.

Subali, B., Kumaidi, N., A., S., & Sumintono, B. (2019). Student achievement based on the use of scientific method in the natural science subject in elementary school. *J. Pendidik. IPA Indones*, *8*(1), 39–51.

Tjabolo, S. A., & Otaya, L. G. (2019). Quality of school exam tests based on item response theory. *Univers. J. Educ. Res., 7* (10), 2156–2164.

Von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., & Khorramdel, L. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ. Princ. Policy Pract*, *26*(4), 466–488.