Research Article

Mathematics Test Development By Item Response Theory Approach And Its Measrument On Elementary School Students

Viktor Pandra[1], Badrun Kartowagiran[2], Sugiman[3]
[1]Yogyakarta State University. Jalan Colombo No. 1, Karangmalang, Yogyakarta 55281, Indonesia. viktorpandra@ymail.com
[2]Yogyakarta State University. Jalan Colombo No. 1, Karangmalang, Yogyakarta 55281, Indonesia.
[3]Yogyakarta State University. Jalan Colombo No. 1, Karangmalang, Yogyakarta 55281, Indonesia.

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: This research is a development and measurement research. This research is conducted to develop a good instrument in measuring elementary school students' skill by IRT approach which based on; 1) validity and reliability of instrument; 2) assumption test; unidimensional, local independence and parameter invariance; 3) characteristics of item test and 4) measuring students' ability. This research is developed by the development model of DjemariMardapi, such as: 1) arranging test specification, 2) writing the test, 3) examining the test, 4) conducting trial test, 5) analyzing the item test, 6) improving the test, 7) assembling the test, 8) implementing the test and 9) interpreting the result of test. The research founding; 1) the range of aiken value developed for instrument test of grade 3 is 0.83-1. While the reliability coefficient of instrument test of math skill in grade 3 is 0.883. the result indicates that the item test developed has good validity and reliability. 2) unidimensional test is fulfilled since the test is proven only measure one dominant dimension i.e., the same skill. The assumption test of local independence is also fulfilled due to the value of covariant among interval of skill is small or close to zero. The calculation result between difficulty level from response includes in high category, therefore the assumption of parameter invariance of skill is fulfilled; 3) based on analysis of the three instruments of fit model result, it fits on 2Pl model, so on the parameter estimation of the item, overall package estimates on 2PL model or parameter b and a (difficulty level and trick). Based on the analysis result of parameter item test in grade 3 indicates that the overall item is on good categories of difficulty level and trick; 4) the measurement of students' skill indicates that the average of students' skill is 0.451 with the maximum skill score of students is 3.185 and the minimum students' skill is -2.282. if it is observed from students' average score, students in the research samples have good average of math skill.

Keywords: Test Development, Mathematics, Item response theory (IRT),Elementary School

1. INTRODUCTION

Trends in Mathematics and Science Study (TIMSS) results (2012) attended by grade VIII Indonesian students in 2011. For mathematics field, Indonesia is ranked 38th with the score of 386 of 42 countries. Indonesia's score is down 11 points from the assessment in 2007. Likewise, the Programme for International Student Assessment (PISA) under the Organization Economic Cooperation and Development (OECD) (2013) conducted a survey on student skill and education systems. Students' skill is assessed in this survey, such as math skill, reading skill and scientific skill (science) that reflects the education system in their each country. The survey results show that the math skill of students in Indonesia ranked 64th out of 65 countries or second from the bottom with a score of 375.

The results of the TIMSS and PISA survey show that Indonesian students' math skill is still low, both in the content dimension and cognitive dimension. Assessment of the dimension of content on domains: numbers, algebra, geometry, data and opportunities, while the assessment of cognitive dimensions on the domain: 1) knowledge, includes facts, concepts and procedures that students should know; 2) application, focusing on students' skill to apply knowledge and understanding of concepts to solve problems or answer questions; 3) reasoning, focusing on solving non-routine problems, complex contexts and performing many problem-solving steps.

In line with the statement above, based on the results of report from the research conducted by the Center for Development of Mathematics Teachers examining in several elementary schools in Indonesia revealed that 51 percent of students have difficulty in the aspect of counting, 50 percent of students have difficulty in mastering concepts, and 49 percent of students have difficulty in solving story problems (PPPG Mathematics Team, 2001: 18). Furthermore, in 2002 based on the results of research from the PPPG Mathematics Team revealed that in some areas of Indonesia, most elementary school students have difficulty in solving story problems and interpreting the story problems into mathematical models (PPPG Mathematics Team, 2002: 71)

There has been no unanimous agreement among mathematicians until today, what it should be called mathematician. According to Hans Freudental (Marsigit, 2013: 10) mathematics is human activity and must be associated with reality. Therefore, when students do learning activity of mathematic, there is mathematical process.

There are two types of matematization, such as: (1) horizontal matematization and (2) vertical matematization. Horizontal matematization process from the real world into mathematical symbols. The process occurs in the student when he/she is faced with real life/situation problem. While vertical matematization is a process occurring within the mathematical system itself; for instance: finding strategy to solve problem, linking relationship between mathematical concepts or applying formulas/ formula findings.

Begle (1979: 6) classifies direct objects in learning mathematics into fact, concept, skill, and principle. Fact is mathematical object which is convention that can be expressed in symbols. A concept is an idea or idea formed by looking at the same character of a set of appropriate object. Skill is a procedure or set of rule used to solve math problems. A principle is a statement of true value, containing two or more concepts and stating the relationship between the concepts.

In order to students have the skill to solve math problem, so the fact, concept, skill and principle are needed. For instance, if students are asked to calculate the area of a flat plane in the form of isosceles triangle, of course they should understand the concept of isosceles triangle, use certain symbol (fact) when constructing a formula for the area of isosceles triangle, have skill in performing calculations of the area of isosceles triangle, and understand the principles in determining and using the formula of the area of the isosceles triangle.

Teacher should facilitate students to learn mathematic through the process of experiencing, therefore students will understand and conceive about the fact, concept, skill, and principle that can be used for problem solving both routine and non-routine. The learning process should provide students with first-person experience to construct knowledge, skill, and ethical attitude (Ramadhan, S., Nasran, S. A., Utomo, H. B., Musyadad, F., &Ishak, S. 2019). Therefore, students will have competence and are able to use or utilize mathematics to solve problem they face in their daily lives.

The learning process in educational unit is organized interactively, inspiring, fun, challenging, motivating learners to participate actively, as well as providing sufficient space for initiative, creativityand independence which are appropriate with the talent, interestand physical and psychological development of learners (Regulation of the Minister of Education and Culture No. 65 of 2013). Learners are encouraged to be able to develop their own knowledge through guidance provided by teachers. This view is based on the assumption that mathematics is the activity of human life (Turmudi, 2008 : 7) or "mathematics as human sense-making and problem solving activity" . In mathematics learning, students should be stimulated to find themselves, conduct their own investigation, prove conjecture themselves, and find out the answers to their friends or teachers' questions.

The material scope and level of competence of learners that should be fulfilled or achieved in an educational unit at a certain level and type of education are formulated in the Content Standards for each subject. Content Standards are criteria regarding the scope of materials and the level of competence to achieve the competence of graduate at a certain level and type of education. The scope of material is formulated based on mandatory content criteria stipulated which is appropriate with the provision of legislation, scientific conceptand characteristic of educational unit and educational program. Furthermore, the level of competence is formulated based on the criteria of the level of learners' development, the qualification of Indonesian competency, and the mastery of tiered competency (Government Regulation No. 32 of 2013).

Core Competency is the translation or operationalization of SKL in the form of quality that should be owned by those that have completed education at a certain educational unit or certain level of education, an overview of the main competencyclassified into aspect of attitude, knowledge, and skills (affective, cognitive, and psychomotor) that should be learned by students for a level of school, class and subjects. Core Competencies should describe the balanced quality between the achievement of hard skills and soft skills (Ministry of Education and Culture, 2013: 5). Core competencyis designed in four interconnected groups, such as with regard to religious attitude (core competency 1), social attitude (competency 2), knowledge (core competency 3) and the application of knowledge (competency 4).

In the curriculum of elementary school mathematics education (Curriculum, 2004), it is mentioned that the effort of improving the quality of education needs to be implemented thoroughly which include aspects of knowledge, skill, attitude and others. The development of these aspects is implemented to improve and develop life-skill through a set of competencies, so that students can survive, adjust and succeed in the future. These skills require systematic, logical, critical thinking skills that can be developed through problem solving in mathematics learning. Therefore, the construction of mathematics problem encourages students to maximize their thinking skill and give flexibility for students to develop problem solving skill based on their experience in daily life.

Students' thinking skill in solving math problem will be reflected in solving math problem. Therefore, the steps to solve math problem tend to be unlimited and vary in order, depending on students' skill in mastering math materials. Math problem tends to be expressed through questions or statements combined with various forms such as story, table, graph and diagram. Aspect of skill measured in math problem include aspect of memory, understanding, application, analysis, syntheses, and evaluation which is appropriate with Bloom's taxonomy.

According to DjemariMardapi (2012: 110) there are nine steps that should be taken in arranging a standardized study test, such as: (1) compiling test specification, (2) writing test, (3) examining test, (4) conducting test, (5) analyzing test item, (6) improving test, (7) assembling test, (8) implementing the test and (9) interpreting test result. Setting the test specification is elaborating the overall characteristics that a test should have. The procedure for drafting test specification include: determining the test objective, arranging the test grid, determining the test form and determining the length of the test.

Hambleton &Swaminathan (1985: 226) state that the process of developing test with item response model, includes: (1) preparation of test specification, (2) preparation of question pool, (3) implementation of test in the field, (4) selection of test question, (5) compilation of norm reference (for norm-referenced tests), (6) specification of pass limit score (for criterion-referenced test), (7) reliability study, (8) validity study and (9) final test production. Meanwhile, according to Shultz & Whitney (2005: 51) the stages of test development include: (1) the preparation of test specification, (2) defining the test domain which include: test context, test construct, dimensionand test pressure, (3) determining the test format, (4) determining the test length, (5) determining the difficulty level test.

The parameter of invariance item has an important consequence, if the item is relatively numerous, the item parameter can be estimated for items that are not answered by the testee. This is known as person free item calibration or free individual item calibration. Wright & Stone (1979) describe that the calibration process could be used in detail on Rasch model. The way conducted isif all items are fit for logistic model 2 parameters, if selected A and B as group 1 and part B and C as group 2. Group 1 has an average score of 0 and a standard deviation of 1, this is same with group 2 if it has an average of 0 and a standard deviation of 1.

The response function of an item contains two parameters such as item parameter and skill parameter. According to Baker (2001: 134) calibration in IRT is the process of determining the parameter of an item and the skill parameter of the item response function. Wells, Subcoviak, &Serlin (2002) state that the calibration process is used to estimate the parameter of the problem grain and observe the skill of the item in distinguishing between latent trait level. Meanwhile, according to Yen & Fitzpatrick (2006: 129) state that calibration is the process of determining the estimation of item parameter and the skill of item response data on IRT. So, calibration is the process of determining the estimation of grain parameters and the parameters of the ability to be known its position in the test instrument.

Until now, there are still a number of problems that are often encountered in schools related to the quality and implementation of assessment activity, especially in the elementary school level. This problem is related to the objective, planning, implementation, result and follow-up of assessment result conducted by both teachers and schools. The problem is based on the result of research conducted by DjemariMardapi, et al. (1999: 45) reveal that there are still many teachers in making test questions not based on the test grid, but tend to only use the questions on the books. Likewise, the results of research conducted by Kumaidi (2005: 5-6) reveal that teacher in compiling test less or not even make the grid first. Many teachers are less in utilizing the test result data to improve the learning process, but it is more used to give the label for students as graduating or not graduating or giving a report number. The form of the number or status of the student is a label given by the teacher which may subsequently have a poor implication.

The result indicates that teacher in preparing testonly used to directly writing the detail of the question without being accompanied by good planning. The planning relate with the determination of behavior aspect or skill tested, determination / selection of essential material, determination of the proportion of cognitive aspects (memory, understanding, application) for each basic competency / indicator, and so on. In addition, teachers do not use the assessment data to find out the extent of their learning success and the strength or weakness of students which need to be responded to make improvements or enrichment.

Another problem isassessment conducted by teachers or schools are often interpreted only for the purpose of giving grades to students, so that the real purpose of assessment is to know how far students have been able to master a basic material / competency taught. Similarly, giving the score for students is often based only on a percentage of a student's correct number of answers on a test without taking into account the weight of each item that builds up that test. As a result, the results of the assessment conducted by the teacher become biased and unable to describe the true competence of the students.

This research conducted try to develop the test to measure the math skill of elementary school students that can be used to identify the level of math skill, measure the development of mathematics skill and develop a profile of the achievement level of the student's math skill to reveal the aspects of the skill tested, whether it is successful or failed to be mastered by the student and strength and weakness of the student.

2. RESEARCH METHOD

This research is a development research with quantitative approach, which aims to produce a product. The product in this research is mathematics skill test instruments of public elementary school students. The product is produced through instrument development procedure.

Instrument development model in the form of test use modification of the Wilson Model (2005: 18) and Order and Antonio Model (1998: 34) with the following steps: (1) initial development of the test, (2) test trial, and (3) broad-scale trial. Initial development consists of: test design and validation by experts' judgement. After the test design is complete, the test is validated by an expert, if there is an item that is not yet reliable, the test is revised first until the test is valid in content (Ramadhan, S., Sumiharsono, R., Mardapi, D., &Prasetyo, Z. K. 2020). Then the instrument is tested on students in grade III elementary school. Based on trial, unfit items were revised and fit items were assembled as fit mathematical tests. This mathematical test is ready to be used for measuring, then continued the broad-scale trial process.

1. Initial Development Test

1.1. Determining the Test Goal

On the stage of initial development test, first that should be done is determining the test goal. This instrument includes summative test due to given at the last final semester. Do, the goal of the test is to know the students' math skill of elementary school.

1.2. Determining the Competency Tested

After the test goal is clear, the next step is chosen the competency tested. This competency is appropriate with the core and basic competency for math subject in grade 3 of elementary school. Based on the core and basic competency, then determined the appropriate indicators.

1.3. Determining the Material Tested

Based on the competency standard, basic competency and indicator, the next is describing Math material of grade 3 of elementary school which is appropriate. An appropriate math material for grade 3 of elementary school include: number, geometry and measurement.

1.4. Arranging the Test Grids

To able to make a good item test, it needs a grid of test. The grid is a matrix containing the specification of test items made. These grids are the guidance of question made, therefore, by the test grids anyone making question, will produce the question and the difficulty level which is relatively same.

1.5. Writing the Item

As it has been stated above, the test grid has crucial role in test development. The test item is made based on the test grid.

1.6. Arranging the Scoring Guideline

The test can be used if it is completed by the guideline of scoring. The scoring guideline is designed to maintain the objectivity of assessment and scoring certainty obtained by the test participant.

1.7. The Content Validity

After the items are compiled in the math skill test in grade III of Elementary School and the scoring guideline is conducted the limited trial. This limited trial was conducted with the aim to find out the readability of the test details. Limited trial results are used as the basis for revision and refinement of the items. Besides limited trial, in order to obtain good instruments, the lattice of instruments, items, and scoring guidelines that have been compiled are subsequently reviewed, and validated. The validation process, in order to meet the requirements in terms of concept, construction and language is used with expert's judgement.

1.8. Item Improvement and Assembly Line of the Test

To make improvement to the test item, qualitative analysis of test quality on the grid, instrument items, and assessment guidelines are conducted first. The first step is to examine aspects, sub-aspect, indicators, and instrument grids. Second, a review of all the test items that have been completed is compiled. Third, all the items that have been compiled are tested on a limited basis. Finally, item improvement is based on limited trial results and expert forum study.

2. Instrument Trial

The trial stage in this case is named limited trial which consist of several ways such as 1) determining the subject trial, 2) implementing the trial and 3) analyzing the trial result.

3. Wide Scale Trial

Wide scale trial has a goal that not only to determine the characteristics of instrument but also determine the individual skill of the respondence. This stage includes: 1) test assembly, 2) wide scale test implementation, 3) result analysis, 4) result interpretation.

4. Design and Trial Subject

The trial will be conducted in grade 3 of elementary school in the area of Lubuklinggau City such as SD Negeri 11 Lubuklinggau, SD Negeri 42 Lubuklinggau, dan SD Negeri 58 Lubuklinggau.

The test subjects in this research are elementary school students since the students are the main users of the product developed in this research. Besides that, students are required to obtain a coefficient of test reliability and wearability of developed tests. The research subjects of elementary school students selected in this study will be used to collect data on students' math skill. The selected elementary school students are grade III, IV, and V. Item instruments developed is 40 items and used to measure grade III. In the limited trial, 160 students are taken and 228 students for measurement.

5. Technique and Instrument of Data Collection

Data collection in this research is done by stratified random sampling technique. The sampling step of the research begins with identifying the strata of schools that are members of the population. Determination of school strata by considering the category of school that is distinguished from excellent and non-excellent school. After the school strata is identified, three schools are randomly selected in each school category. All grade III students in the selected schools are the samplesof research.

The instrument or data collection tool used in this research is a test, a test of mathematics skill developed by researcher based on Core Competency / Basic Competency / teaching material in elementary school. Determination of Core Competency, Basic Competency and indicator tested are conducted through Group Discussion Forum (FGD). FGD participants are teacher, subject teacher, study expert (mathematics lecturer) and researcher. FGD participants choose Core Competency, Basic Competency and indicator that are reliable or important to be tested at summative tests (end of semester).

6. Data Analyzing Technique

The respondent's answer sheet is corrected and discrete by the assessor. Assessors are elementary school math teachers that have attended the training. The training was conducted twice, namely: (1) equalling understanding of the contents of the test details, and (2) equalling the understanding of the way of scoring. The data of test result are analysed quantitatively. Analysing the items use customized scales. Data has been already in the format and analysed by using BILOG-MG program. Based on the analysis of the test results, obtained the parameters of the test item, so that it can be done the improvement of the details of the question that is deemed necessary. Proof of validity based on internal structure can be verified by CTT and IRT (Vendramini&Silvi, 2011:1). Therefore, based on the analysis of test results, it is found out: (1) the details of the problem that are not fit and (2) the coefficient of reliability. If the test item is not yet eligible, it is corrected. However, if all items in the instrument in the form of a test device are met, mathematic test device can be used to test mathematic skill. Data analysis techniques consist of several aspects, namely: a) Reliability, b) Goodness of fit, c) Difficulty Level, d) Information Function and SEM.

6.1. Reliability

The estimation of reliability in this research use the formula of Cronbach-Alpha. The reliability estimation is supported by SPSS 22 program.

6.2. Goodness of Fit

Data of trial result which is the form of learners or respondents' answer are then conducted a goodness of fit analysis of the model. Fit testing is performed on the item analysis which is scoreddichotomous. Test of goodness of fit for the overall test as well as each item use the BILOG-MG program.

6.3. Difficulty level*(b)*

The second characteristic use the difficulty level or index difficulty by utilize the BILOG-MG program to obtain index difficulty or difficulty level *(b)*. The item can be stated as good item if the index of difficulty is more than -2.0 or less than 2.0 which is able to be stated with ($-2,0 < b < 2,0$).

6.4. Information Function and SEM

Based on the analysis with Pascale, obtained the information function and standard error of measurement (SEM). Based on the information function and SEM, this test is suitable for learners having low, medium and high skill(θ).

3. RESULT AND DISCUSSION

1. Development Result of Initial Product

One of the educational problems related with the quality of education is the low mastery of students toward the competency, as a result of inadequate assessment. The assessment system is not optimal because: (1) the quality of test made by teacher is still inadequate, (2) the monitoring of the testing network in the area has not been implemented properly, (3) the reporting of exam results has not been optimal and (4) the utilization of exam results has not been done optimally.

Based on the description above, this research try to develop math skill test for elementary school students in grades III, IV, and V. Development of math skill test refers to Core Competency and Basic Competency based on Curriculum 2013. The test is used to: (1) identify the math level of elementary school students, measure the progression of elementary school students' math skill and profile the achievement level of the student's math skill.

1.1. Content Validity

Validity is classified into three types, such as: (1) validity of content, (2) validity of criteria (criterion-related)and (3) validity of construct (Nunnally, 1978, Allen & Yen, 1979, Fernandes, 1984, Woolfolk &McCane, 1984, Kerlinger, 1986, and Lawrence, 1994). This validity can be found out through the analysis of the contents of the test and empirical analysis of the test score of grain response data (Lissitz&Samuelsen, 2007). The validity of the contents of an instrument is defined as to what extent the items in the instrument represent the components in the entire content of the object to be measured and to what extent they reflect the characteristics of behaviour to be measured (Nunnally, 1978; Fernandes, 1984).

The validity of the content is determined by using expert agreement. Expert agreement of the field of study or often referred to a measured domain determines the level of content validity related (HeriRetnowati). This case is caused by the measurement instrument, such as test or questionnaire is proven valid only if the expert believes that the instrument is able to measure the mastery of the skill defined in the measured domain. Analysing the validity of the content use the aiken formula.

Aiken formulate the formula Aiken’s V to calculate the content validity coefficient which is based on the result of assessment from the experts’ panel as much as n of people toward an item of to the extent of the item represent the construct measured.

The instrument can be stated valid if the experts believe that the instrument measure the things which will be measured. Experts’ judgement give the scoring used that will be used to prove the content validity toward the number of instruments in this research. The instrument that will be validated are as follow:

Table 1. The Result of Instrument Validity Developed

Item	R1	R2	R3	S1	S2	S3	Number S	Value V
1	5	5	5	4	4	4	12	1.00
2	5	5	5	4	4	4	12	1.00
3	5	5	5	4	4	4	12	1.00
4	5	5	4	4	4	3	11	0.92
5	5	5	5	4	4	4	12	1.00
6	5	4	5	4	3	4	11	0.92
7	5	5	4	4	4	3	11	0.92
8	5	5	5	4	4	4	12	1.00
9	5	5	5	4	4	4	12	1.00
10	4	5	5	3	4	4	11	0.92
11	5	5	5	4	4	4	12	1.00
12	5	5	5	4	4	4	12	1.00
13	4	5	5	3	4	4	11	0.92
14	5	5	5	4	4	4	12	1.00
15	5	5	5	4	4	4	12	1.00

Item	R1	R2	R3	S1	S2	S3	Number S	Value V
16	5	5	4	4	4	3	11	0.92
17	5	5	5	4	4	4	12	1.00
18	5	5	5	4	4	4	12	1.00
19	5	5	5	4	4	4	12	1.00
20	5	5	5	4	4	4	12	1.00
21	5	5	5	4	4	4	12	1.00
22	5	5	5	4	4	4	12	1.00
23	5	5	5	4	4	4	12	1.00
24	5	5	5	4	4	4	12	1.00
25	5	5	5	4	4	4	12	1.00
26	5	5	5	4	4	4	12	1.00
27	5	4	5	4	3	4	11	0.92
28	5	5	5	4	4	4	12	1.00
29	5	5	5	4	4	4	12	1.00
30	5	5	5	4	4	4	12	1.00
31	5	5	5	4	4	4	12	1.00
32	5	4	5	4	3	4	11	0.92
33	5	5	5	4	4	4	12	1.00
34	5	5	5	4	4	4	12	1.00
35	5	5	5	4	4	4	12	1.00
36	5	5	5	4	4	4	12	1.00
37	5	5	5	4	4	4	12	1.00
38	5	5	5	4	4	4	12	1.00
39	5	5	5	4	4	4	12	1.00
40	5	4	5	4	3	4	11	0.92

Table 1 is the validation result of instrument which use the index of Aiken V. based on the data above, the range of aiken value for the instrument is 0.83-1. While based on aiken table, if the number of items is 40, consist of 5 criteria and there are 3 raters, so the minimal limit accepted is 0.92. based on the data, it can be stated that all the items are proven valid reviewed from the content of validity except item number 17 and 19, therefore need to be revised again.

1.2. Reliability

The reliability of a test is generally expressed numerically in a coefficient of $-1.00 \leq \rho \leq +1.00$ (Retnawati, 2016). Mahrens & Lehman (1973) state that although there is no general agreement, it is widely accepted that for test used to make decisions on individual students should have a minimum reliability coefficient of 0.85. The estimated reliability of the research used the Cronbach-alpha formula and was analysed with the support of the SPSS 22 program. The estimated reliability results on three instruments are presented below

Table 2. The Result of Reliability Estimation

Instrument	Cronbach-alpha	Number of Item
Instrument of grade III	0,883	40

Table 2 is reliability estimation result in research instrument developed. This research estimate the reliability on the instrument of grade 3 where the instrument consist of 40 items of question.

Table 2 above explain that the instrument of grade 3 has coefficient of reliability which is 0.883. The result of coefficient estimation of reliability show that the instrument develop is reliable to be used in measuring the students' math skill of elementary school grade 3.

2. Instrument Trial Result

The trial instrument result is begun with assumption test. Assumption test is a precondition test to find out whether the result of research is reliable to be conducted to the next test step or not. The precondition test in this research consist of unidimensional test, independence of local and parameter invariance.

2.1. Unidimensional Test

Assumption test which should be fulfilled is that every item of test only measures one skill. One of the ways to test this assumption is by analysis factor producing KMO, Eigen value and variant that can be explained and the component of factor. Analysis of exploratory factor is conducted by the support of SPSS 22. The result of factor analysis on the three instruments developed can be presented on the table 6.

Table 3. Test Result of KMO

	Instrument of grade III
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0,819
Bartlett's Test of Sphericity (Approx. Chi-Square)	1909,788
Df	780
Sig	0,000

Table 3 is the analysis test result of KMO on the instrument of grade 3. Table 3 show that the value of KMO on the instrument of grade 3 is 0.819. the result is bigger than 0.50 which means that the three of trial samples used in this instrument is stated enough. The matrix can be conduct factor analysis if the value of KMO is bigger than 0.5.

Table 4. Eigen value of Math Test Instrument of Grade III

Component	Grade III		
	Initial Eigenvalues		
	Total	% of variance	Cumulative %
1	**7.331**	**18.327**	**18.327**
2	**1.679**	**4.198**	**22.525**
3	**1.610**	**4.025**	**26.550**
4	**1.470**	**3.676**	**30.226**
5	**1.399**	**3.496**	**33.723**
6	**1.316**	**3.290**	**37.013**
7	**1.280**	**3.199**	**40.212**
8	**1.248**	**3.119**	**43.331**
9	**1.207**	**3.018**	**46.349**
10	**1.168**	**2.920**	**49.269**
11	**1.1**	**2.873**	**52.142**

Component	Grade III		
	Initial Eigenvalues		
	Total	% of variance	Cumulative %
	49		
12	**1.116**	**2.789**	**54.931**
13	**1.030**	**2.574**	**57.505**
14	.978	2.445	59.950
15	.969	2.422	62.372
16	.932	2.331	64.703
17	.902	2.255	66.958
18	.865	2.164	69.121
19	.818	2.045	71.166
20	.803	2.007	73.173
21	.767	1.917	75.089
22	.756	1.890	76.979
23	.736	1.840	78.819
24	.698	1.745	80.564
25	.671	1.678	82.242
26	.651	1.629	83.870
27	.610	1.524	85.395
28	.592	1.479	86.874
29	.575	1.439	88.312
30	.550	1.376	89.688
31	.536	1.339	91.027
32	.491	1.227	92.254
33	.470	1.174	93.428
34	.455	1.138	94.567
35	.439	1.098	95.665
36	.405	1.012	96.677

Component	Grade III		
	Initial Eigenvalues		
	Total	% of variance	Cumulative %
37	.389	.972	97.649
38	.354	.885	98.534
39	.303	.757	99.291
40	.283	.709	100.000

Table 4 is the eigen value of math test instrument of grade III. The number factors formed can be view from eigen value >1, which means that the factor used as an indicator (Wagiran, 2014: 302).On the instrument of grade III show that from 40 item of questions form 13 factors, where the factor 1 is dominant factor with the eigen value is 7.331. factor 1 as dominant factor is a factor having the highest eigen value compared with other factors, therefore, it can be stated that the instrument developed is unidimensional.

Dimension that is measured in a data can be proved on the result of scree plot, i.e., the amount of steep. The number of steeps show the number of dimension or factor and the change of eigen value do not show the existence of dimension (Retnawati, 2016: 142). Therefore, unidimensional also can be viewed from the result of scree plot formed. The test is stated unidimensional when the component 1 and 2 in scree plot have the distance which is far enough (Furr& Bacharach, 2008: 74).

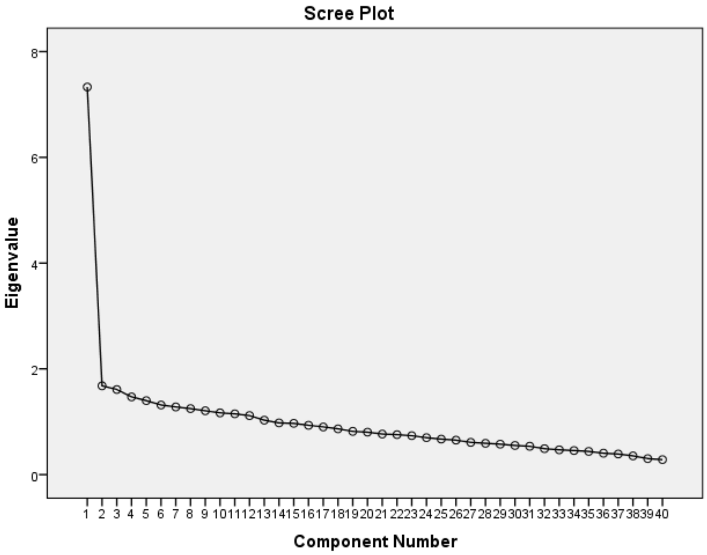

Figure 1. Scree plot exploratory analysis result of analysis factor

Figure 1 is scree plot exploratory analysis result of analysis factor from the instrument of grade III. Figure 1 show that all instruments on component 1 have far range with component 2, while component 2 to component 3 has really close range. This case indicates that there is one dominant factor and other factors give a great contribution toward the variant that can be explained. Based on the scree plot above, all instrument developed in this research is considered unidimensional.

2.2. Local Independence Test

Assumption of local independence isthe requirement that should also be fulfilled if using IRT analysis. This assumption test aims to figurewhether the student's skill is independent toward the item, which means that the student's answer to one item will not affect the answer to the other item. The assumption test of local independence can be proven automatically after proven by the unidimensional of participants' response data to the test (Retnawati, 2014: 7). However, local independence assumption tests can also be proven through a covariant matrix based on the skill of students that classified into several groups. This assumption is fulfilled if the covariance value between the skill interval is small or close to zero. Therefore, if the covariant value is close to zero, then it can be concluded that it fulfils the assumption of local independence.

Table 5. Test Result of Local Independence

	K 1	K 2	K 3	K 4	K 5	K 6	K 7	K 8	K 9	K 10
K 1	0.3690									
K 2	0.1011	0.0460								
K 3	0.0625	0.0243	0.0158							
K 4	0.0475	0.0209	0.0111	0.0104						
K 5	0.0354	0.0141	0.0082	0.0062	0.0060					
K 6	0.0418	0.0173	0.0103	0.0076	0.0067	0.0082				
K 7	0.0433	0.0183	0.0095	0.0079	0.0051	0.0064	0.0093			
K 8	0.0322	0.0161	0.0090	0.0079	0.0052	0.0061	0.0052	0.0081		
K 9	0.0596	0.0211	0.0128	0.0103	0.0072	0.0082	0.0075	0.0094	0.0145	
K 10	0.1654	0.0750	0.0431	0.0330	0.0255	0.0304	0.0296	0.0279	0.0370	0.6266

Table 5 is covariant matrix based on students' skill of grade III. The table indicates that matrix value of variant-covariant among groups of students' skills. Based on the analysis result, it is found out that the variant covariant value among groups of intervals of students' skill that form diagonal line is small even close to zero. Therefore, there is no correlation and it can be concluded that local independence has been fulfilled.

2.3. Parameter Invariance

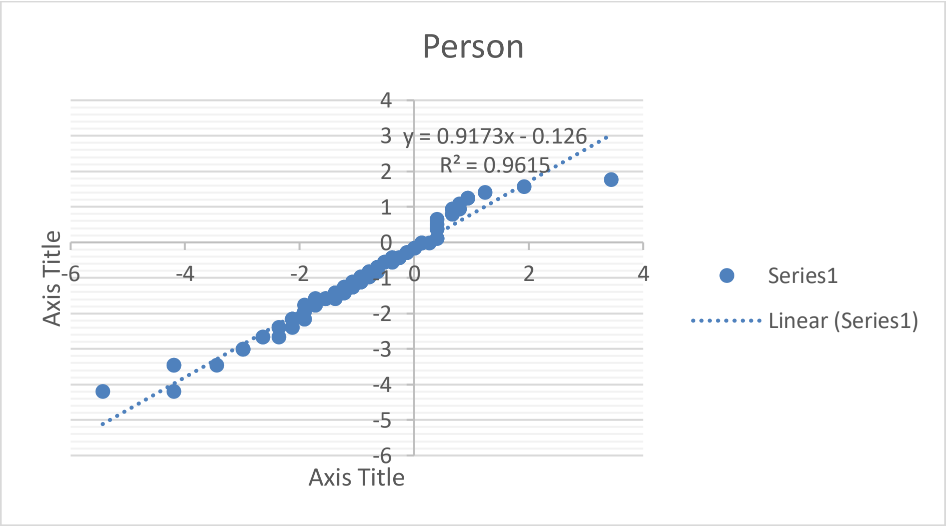

Figure 2. Parameter Invariance of Respondent

Figure 2 is analysis result of parameter invariance in grade III. The calculation result of correlation between the difficulty level of the response in grade III include in high category which is 0.9615. the result of parameter estimation of each sample's skill then made scree plot and correlated. If the correlation is positive and high, the invariance assumption of skill parameter is fulfilled (Retnawati, 2014: 9).

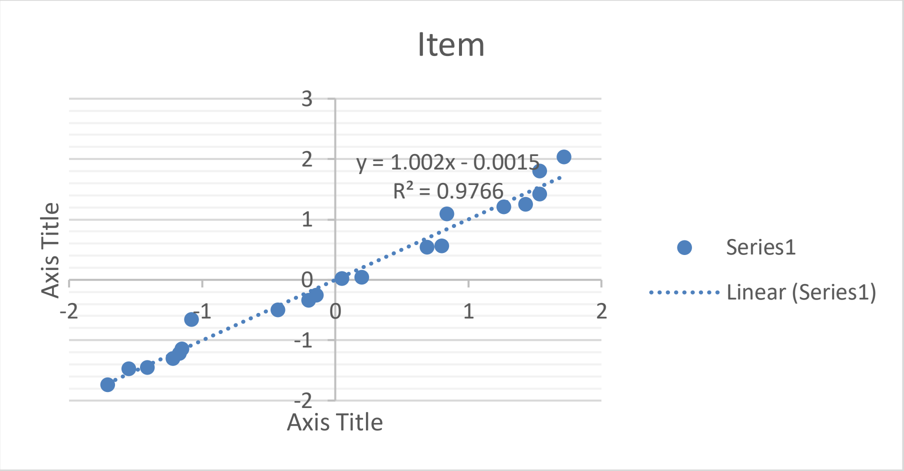

Figure 3. Parameter Invariance item

Figure 3 is parameter invariance analysis result of grade III. The result of scree plot of grade III explains that the estimation result is really close to straight line and the correlation value is 0.9766 which includes in very high category. The parameter invariance assumption of skill can be concluded that it has been fulfilled.

2.4. Test f Goodness of Fit

The three assumptions for IRT analysis have been fulfilled well, so it can be conducted the goodness fit test model for test analysis that has been developed. Goodness of fit tests model for 1-PL, 2-PL, or 3-PL were performed by comparing the valueof x^2.. The probability value of each item shouldfulfil p>0.05, otherwise revision isconducted before the instrument testing is conducted. The goodness of fit test was analysed by using the support of MG Bilog program. The following table is the result of goodness of fit model analysis that has been done.

Table 6.Test Result of Goodness of Fit Model

Item	Parameter Logistic			Explanation		
	PL 1	PL 2	PL 3	PL 1	PL 2	PL 3
1	0.0455	0.1098	0.0006	Suitable	Suitable	Unsuitable
2	0.9991	0.9906	0.1384	Suitable	Suitable	Suitable
3	0.7606	0.9691	0.1840	Suitable	Suitable	Suitable
4	0.4290	0.7490	0.0867	Suitable	Suitable	Suitable
5	0.7627	0.9889	0.3955	Suitable	Suitable	Suitable
6	0.6888	0.4187	0.0462	Suitable	Suitable	Unsuitable
7	0.7338	0.9575	0.2078	Suitable	Suitable	Suitable
8	0.9916	0.9925	0.7401	Suitable	Suitable	Suitable
9	0.9804	0.9970	0.6999	Suitable	Suitable	Suitable
10	0.3342	0.1815	0.0738	Suitable	Suitable	Suitable
11	0.4900	0.9338	0.1332	Suitable	Suitable	Suitable
1	0.29	0.62	-	Suit	Suit	Unsuitabl

Item	Parameter Logistic			Explanation		
	PL 1	PL 2	PL 3	PL 1	PL 2	PL 3
2	01	63		able	able	e
13	0.2682	0.9118	-	Suitable	Suitable	Unsuitable
14	0.9542	0.9991	-	Suitable	Suitable	Unsuitable
15	0.5895	0.2735	-	Suitable	Suitable	Unsuitable
16	0.3546	0.7726	-	Suitable	Suitable	Unsuitable
17	0.4206	0.8744	-	Suitable	Suitable	Unsuitable
18	0.9935	0.1387	-	Suitable	Suitable	Unsuitable
19	0.8682	0.9637	-	Suitable	Suitable	Unsuitable
20	0.8543	0.9603	-	Suitable	Suitable	Unsuitable
21	0.0086	0.0054	-	Suitable	Suitable	Unsuitable
22	0.4460	0.3579	-	Suitable	Suitable	Unsuitable
23	0.1722	0.7061	-	Suitable	Suitable	Unsuitable
24	0.9513	0.9672	-	Suitable	Suitable	Unsuitable
25	0.4165	0.9915	-	Suitable	Suitable	Unsuitable
26	0.2776	0.1337	-	Suitable	Suitable	Unsuitable
27	0.6518	0.8672	-	Suitable	Suitable	Unsuitable
28	0.4534	0.9031	-	Suitable	Suitable	Unsuitable
29	0.9271	0.6739	-	Suitable	Suitable	Unsuitable
30	0.7118	0.8679	-	Suitable	Suitable	Unsuitable
31	0.9042	0.9094	-	Suitable	Suitable	Unsuitable
32	0.9070	0.5563	-	Suitable	Suitable	Unsuitable
33	0.7761	0.9401	-	Suitable	Suitable	Unsuitable
34	0.7907	0.8596	-	Suitable	Suitable	Unsuitable
35	0.2262	0.8114	-	Suitable	Suitable	Unsuitable
36	0.0111	0.4273	-	Suitable	Suitable	Unsuitable
37	0.9789	0.5201	-	Suitable	Suitable	Unsuitable
3	0.50	0.80	-	Suit	Suit	Unsuitabl

Item	Parameter Logistic			Explanation		
	PL 1	PL 2	PL 3	PL 1	PL 2	PL 3
8	76	56		able	able	e
39	0.6887	0.9887	-	Suitable	Suitable	Unsuitable
40	0.8618	0.5762	-	Suitable	Suitable	Unsuitable
Total of good item				40	40	31

Table 6 is the result of goodness of fit model. Table 6 show that the number of items which are suitable for model 1PL and 2PL is 40 items or the overall items are suitable for the models. Meanwhile, on the model of fit item on model 3PL is 31 items. The result on goodness of fit test indicates that the most model which are suitable are model 1PL and 2PL. Based on the case, the model used in this research is model 2PL.

2.5. Parameter Estimation of Item Question

The analysis used to figure out the characteristics of a good item is by using 1 PL model. Items that fit the model with 2 PL are then reanalysed to figure out the characteristics of the item. The criteria for a good item according to model 2 PL are based on the different trick (*ai*) and difficulty level of item (*bi*). Theindex of different trick of item can be stated good if it is between 0-2. Besides that, an item can be stated good if the index of difficulty level range between -2 to +2 (Hambleton &Swaminathan, 1985: 107). The following is the parameter estimation result of item question developed.

Table 7. Estimation Result of Item Parameter

Item	Parameter		Category		Explanation
	Difficulty Level (*bi*)	Difference Trick (*ai*)	Difficulty Level (*bi*)	Difference Trick (*ai*)	
1	-0.615	0.824	Good	Good	Accepted
2	-0.295	0.628	Good	Good	Accepted
3	-1.081	0.702	Good	Good	Accepted
4	-0.837	0.744	Good	Good	Accepted
5	-1.248	0.872	Good	Good	Accepted
6	0.233	0.626	Good	Good	Accepted
7	-0.643	0.811	Good	Good	Accepted
8	-0.522	0.659	Good	Good	Accepted
9	-0.075	0.665	Good	Good	Accepted
10	-1.199	0.478	Good	Good	Accepted
11	-1.392	0.673	Good	Good	Accepted
12	-1.543	0.490	Good	Good	Accepted
13	-1.634	0.796	Good	Good	Accepted

Item	Parameter		Category		Explanation
	Difficulty Level (*bi*)	Difference Trick (*ai*)	Difficulty Level (*bi*)	Difference Trick (*ai*)	
14	-0.479	0.543	Good	Good	Accepted
15	-1.107	0.595	Good	Good	Accepted
16	-1.330	0.885	Good	Good	Accepted
17	-1.537	0.671	Good	Good	Accepted
18	-0.778	0.556	Good	Good	Accepted
19	0.104	0.621	Good	Good	Accepted
20	0.310	0.617	Good	Good	Accepted
21	0.268	0.594	Good	Good	Accepted
22	-0.224	0.705	Good	Good	Accepted
23	0.001	0.527	Good	Good	Accepted
24	0.358	0.805	Good	Good	Accepted
25	-1.665	0.706	Good	Good	Accepted
26	-0.133	0.761	Good	Good	Accepted
27	-0.023	0.702	Good	Good	Accepted
28	0.509	0.818	Good	Good	Accepted
29	-0.036	0.646	Good	Good	Accepted
30	0.148	0.689	Good	Good	Accepted
31	-0.794	0.682	Good	Good	Accepted
32	-0.338	0.567	Good	Good	Accepted
33	-0.417	0.568	Good	Good	Accepted
34	-1.263	0.747	Good	Good	Accepted
35	-0.682	0.595	Good	Good	Accepted
36	-1.397	0.459	Good	Good	Accepted
37	-1.542	0.760	Good	Good	Accepted
38	-1.103	0.627	Good	Good	Accepted
39	0.187	0.677	Good	Good	Accepte

Item	Parameter		Category		Explanation
	Difficulty Level (*bi*)	Difference Trick (*ai*)	Difficulty Level (*bi*)	Difference Trick (*ai*)	
					d
40	0.191	0.721	Good	Good	Accepted

Based on the analysis result of limited trial analysis as presented on the table 7, there are 40 items of question that viewed from the parameter of different trick (*ai*) and difficulty level (*bi*). Based on analysis result, table 7 indicates that the overall is fulfilled good criteria both parameter of difficulty level (*bi*) and different trick (*ai*). On the parameter of difficulty level show the maximum value which is 0.509 and the minimum value is -1.665, while the average of difficulty level is -0.590. Theparameter of different trick shows the maximum value which is 0.885 and the minimum value is 0.459, while the average value is 0.670. based on those results, it can be concluded that the overall items on the parameter of difficulty level and different trick include in good category and ready to be used to conducted the process of measurement.

2.6. Information function(IF) dan standard error measurement (SEM)

Information function is used to reveal the latent skill which measure by using the test through item contribution. Information function of test is also the number of function of each item. Information function is inversely proportional with measurement error or standard error measurement. The value of information function of test instrument will be high if the items of test arrangement have high information function. The following is the curve of the relation between information function and error measurement on each class. The following is IF and SEM analysis result.

Table 8. Analysis Result of IF and SEM

□	**Total of InformationFunction**	**SEM**
-3	1,062	0,971
-2,75	1,463	0,827
-2,5	2,075	0,694
-2,25	2,856	0,592
-2	3,781	0,514
-1,75	4,786	0,457
-1,5	5,771	0,416
-1,25	6,630	0,388
-1	7,279	0,371
-0,75	7,668	0,361
-0,5	7,768	0,359
-0,25	7,561	0,364
0	7,043	0,377
0,25	6,249	0,400
0,5	5,263	0,436
0,75	4,208	0,487
1	3,207	0,558
1,25	2,346	0,653
1,5	1,661	0,776
1,75	1,148	0,933
2	0,779	1,133
2,25	0,523	1,383
2,5	0,348	1,696
2,75	0,230	2,086
3	0,151	2,571

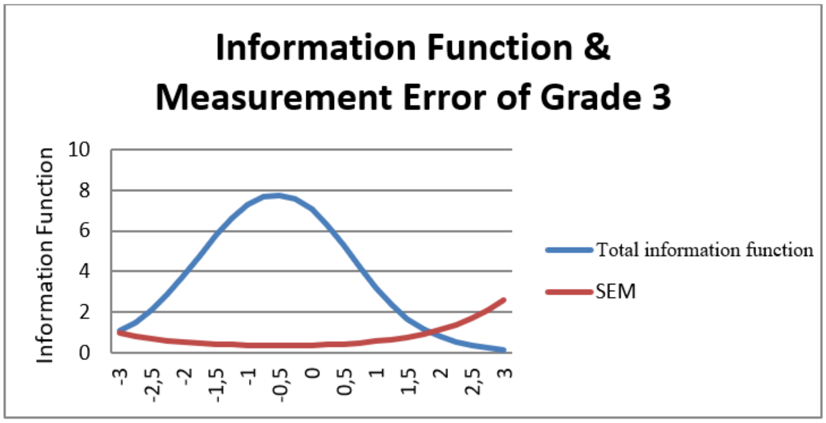

Figure 4. Curve of IF and SEM

Figure 8 and table 4 above are the analysis result of information function and error measurement in grade 3. The curve of information function can be viewed in blue line and the curve of error measurement can be viewed in red line. The figure and table above indicate that the test conducted give the maximum information which is 7.768 and the minimum error measurement is 0.359. Based on the figure show that the test instrument is able to measure students' skill and the number of skill (theta) is -3 to 1.75.

3. Students' Skill Measurement

Students' skill in this research is viewed based on the score logit with very high, high, medium, low and very low category. The number of students samples in grade 3 is 228 students. The following is students' skill category in grade 3.

Table 8. Interval of Skill Category

Interval of Skill	Category
1,819 < X	Very High
0,907 < X ≤ 1,819	High
-0,004 < X ≤ 0,907	Medium
-0,915 < X ≤ -0,004	Low
X ≤ -0,915	Very Low

Students' skill descriptive is an analysis result about students' skill based on the score logit. Analusis of students' skill in this research use Bilog-MG softwere. The following is the analysis result of students' skill in each class.

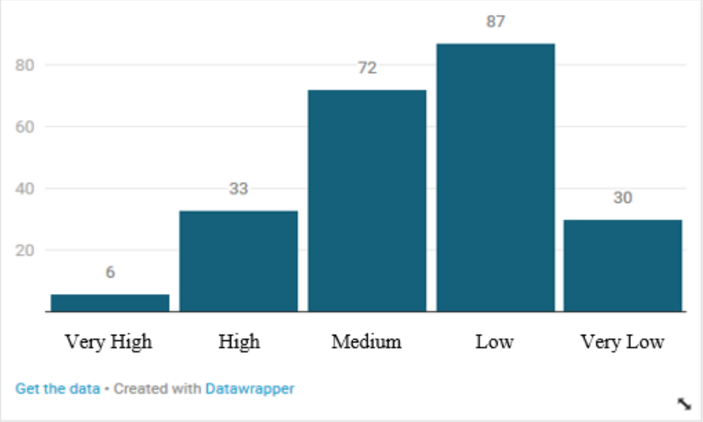

Figure 5. Distribution of Students' Math Skill of Grade 3

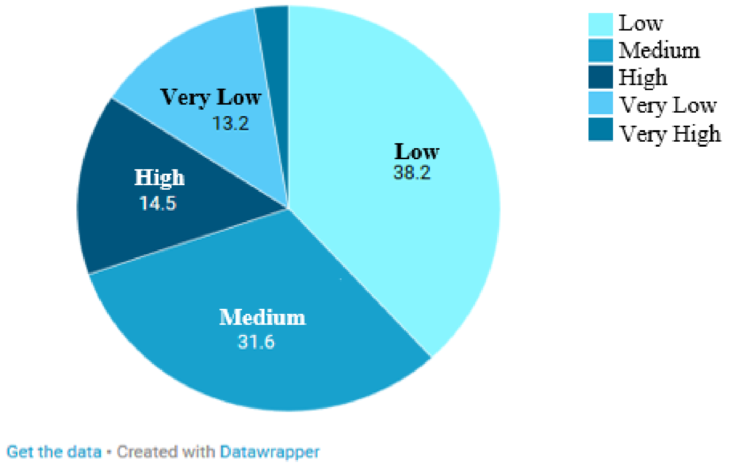

Figure 6. Distribution of Students' Math Skill of Grade 3

Figure 5 and 6 are the analysis result of students' skill by using Bilog-MG program, then the result of skill is visualized into bar chat to view its data distribution and changed into pie chat to view its percentage. Figure 5 and 6 show the number of students in very high category is 6 students or 2.6 percent, in high category is 33 students or 14.5 percent, in medium category is 31.6 percent, in low category is 87 students or 38.2 percent and in very low category is 30 students or 13.2 percent.

The result data of skill analysis, then viewed based on skill maximum, minimum and average score. This case is conducted to find out the descriptive achievement of students' skill based on score logit. The following is analysis result that has been conducted.

Table 9. Data Descriptive of Skill

	Score
Maximum Skill	3,185
Minimum Skill	-2,282
Average Skill	0,451

Table 9 is the result of descriptive analysis of skill data that has been conducted. The table 9 explain that students' skill of maximum score is 3.185, while the minimum score is -2.282. the average score of students' skills is 0.451.

If we observe students' skill distribution on table 5. It can be concluded that students' distribution, students' skill tendency is on medium, low and very low category. However, if it is observed on table 9, students' skill includes in very good category, this case is proven by the average score of students is 0.451. This case explains that the probability of students in doing the test item with the maximum average of difficulty level is 0.451.the difficulty level is on medium category since the difficulty level of medium categoryis between -2 to 2.

If it is observed from the maximum score of skill on the table 9, show that the score is 3.185. the score indicates that students are able to answer rightly the item of test with the characteristics of difficulty level is 3.185. the case show that students are in category having very high skill due to students are able to do the items with high difficulty level. Item with high difficulty level is the item having > 2 of score logit.

If observed from students; minimum score on the table 9 show that the score is -2.282. the score explains that the opportunity of students answers the question rightly only on the item question with the characteristic of maximum difficulty level which is -2.282. the case indicates that students obtaining the score are only able to do the item test with the characteristics -2.282. if observed from the characteristic of question on the level of difficulty, the students include in low category since the students are only able to do the test on the characteristic -2.282.

The exposure above conclude that there is a gap which is far enough between students with medium average skill and students with high skill. Based on the result and discussion above, it explains that the average of students' skill is on medium category and it can be concluded that students in measurement sample have a good mathematics skill.

4. CONCLUSION

Based on the result of development and discussion toward the instrument development to measure students' mathematics skill in elementary school, it can be concluded as follow:

1. The range of aiken value for the instrument test of grade 3 is 0.83-1. While based on aiken table if the number of items is 40, which consist of 5 criteria and 3 raters, the minimum range accepted is 0.92. based on the data it can stated that all items are proven valid based on the content validity. Based on the analysis result show that coefficient reliability of instrument test of math skill in grade 3 is 0.883, so it can be concluded that all instruments developed can be stated reliable.
2. The result of precondition test show that unidimensional test is fulfilled due to the test is proven only measure one dominant dimension such as the same skill. Local Independence assumption test is also fulfilled since the covariant value among intervals of skill is small or close to zero. The calculation result of correlation between the difficulty level of response include in high category, therefore the parameter invariance assumption of skill is fulfilled.
3. Based on the analysis of three instruments, the result of goodness of fit is suitable on the model 2PL, so in parameter estimation of item of all packages estimate on 2Pl model or parameter b and a (difficulty level and difference trick). Based on the parameter analysis result of item test grade 3 show that the overall items is on the categories of good difficulty level and difference trick. The case indicates that the overall items is accepted and reliable to used to measure the development of students' math skill of elementary school.
4. Measurement of students' skill show that the average of students' skill is 0.451 with the maximum students' skill score is 3.185 and the minimum score is -2.282. if observed from the average students' score, so the students in the research sample have good average math skill.

References

1. Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory.* Belmont, CA: Wadsworth, MC.
2. Begle, E.G. (1979). *Critical variable in mathematics education*. Wasingthon, D.C.: NCTM.
3. Depdikbud. (2013). *Peraturan Menteri Nomor 65 Tahun 2013, tentang Standar Proses Pendidikan Dasar dan Menengah*.
4. Djemari Mardapi. (1999). *Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional*. Pidato Pengukuhan Guru Besar. Yogyakarta: Universitas Negeri Yogyakarta.
5. Fernandes, H. J. X. (1984). *Evaluation of educational program*. Jakarta: National Education Planning, Evaluating and Curriculum Development.
6. Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory.* Boston, MA: Kluwer Inc.
7. Hambleton, R.K., Swaminathan, H., & Rogers, HJ. (1991). *Fundamental of item response theory.* Newbury Park, CA: Sage Publication Inc.
8. Kumaidi. (2004). *Sistem asesmen untuk menunjang kualitas pembelajaran*. Jurnal pembelajaran, 27, 93-106.
9. Lissitz, W. & Samuelsen, K. (2007). *Further clarification regarding validity and eduction. Educational Researcher*, Vol. 36, No. 8, pp. 482-484.
10. Mardapi, Djemari (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Litera.
11. Marsigit (2013). *Pendidikan karakter melalui pembelajaran matematika*. Pidato pengukuhan guru besar Universitas Negeri Yogyakarta, disampaikan di depan rapat terbuka senat Universitas Negeri Yogyakarta.
12. Nunnally, J. (1978). *Psychometric theory (2nd ed.).* New York: McGraw Hill.
13. OECD. 2013. *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing.
14. Ramadhan, S., Sumiharsono, R., Mardapi, D., &Prasetyo, Z. K. (2020). The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis. *International Journal of Instruction*, *13*(2), 507-518.
15. Ramadhan, S., Nasran, S. A., Utomo, H. B., Musyadad, F., &Ishak, S. (2019). The implementation of generalisability theory on physics teachers' competency assessment instruments development. *International Journal of Scientific and Technology Research*, *8*(7), 333-337.
16. Retnawati, Heri (2014). *Teori respons butir dan penerapannya*. Yogyakarta: Parama Publishing.

17. Retnawati, Heri. (2014). *Membuktikan Validitas Instrumen dalam Pengukuran*. Diambil dari: http://evaluation-edu.com/wp-content/uploads/2014/10/2-Validitas-heri-Retnawati-uny.pdf
18. Retnawati, Heri. (2016). Validitas*, Reliabilitas, & Karakteritik Butir*. Yogyakarta: Parama *Publishing*
19. Woolfolk, A. E. & McCune, L. N. (1984). *Educational psychology for teachers.* Englewood Cliffs, NJ.: Prentice Hall, In.