Mathematics and Statistics 9(2): 106-111, 2021
DOI: 10.13189/ms.2021.090204 http://www.hrpub.org

Instrument Test Development of Mathematics Skill on Elementary School

Viktor Pandra[1,*], Badrun Kartowagiran[2], Sugiman[3]

[1]Doctoral Program of Educational Research and Evaluation Department, Yogyakarta State University, Yogyakarta, Indonesia
[2]Educational Research and Evaluation Department, Yogyakarta State University, Yogyakarta, Indonesia
[3]Study Program of Mathematics Education, Faculty of Mathematics and Natural Sciences, Yogyakarta State University, Yogyakarta, Indonesia

Received December 14, 2020; Revised February 20, 2021; Accepted March 12, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Viktor Pandra, Badrun Kartowagiran, Sugiman , "Instrument Test Development of Mathematics Skill on Elementary School," Mathematics and Statistics, Vol. 9, No. 2, pp. 106 - 111, 2021. DOI: 10.13189/ms.2021.090204.*

(b): *Viktor Pandra, Badrun Kartowagiran, Sugiman (2021). Instrument Test Development of Mathematics Skill on Elementary School. Mathematics and Statistics, 9(2), 106 - 111. DOI: 10.13189/ms.2021.090204.*

Abstract The aims of this research are: 1) producing the test instrument of mathematics skill on elementary school which is valid and reliable, 2) finding out the characteristics of the test instrument of mathematics skill on elementary school. The instrument test development in this research uses the development model of Wilson, Oriondo and Antonio which is modified. The number of testing sample in this research is 160 students in each class. This research results: 1) the validity index of aiken v is 0.979 in grade IV and 0.988 in grade V. The coefficient of instrument skill in class IV and V are 0.883 and 0.954. 2) the compatibility model in this research is it is suitable for 1PL model or parameter b (difficulty level). The result of parameter analysis of test item in class IV and V, shows that the overall item is in good category which is between -2 to 2. The case indicates that the overall item is accepted and reliable to be used for measuring the development of mathematics skill of elementary school students.

Keywords Development Test, Characteristics of Instrument, Mathematics, Elementary School

1. Introduction

Elementary school is an educational institution which has a function to impart the basic ability and skill for the necessity of the lesson continuity on higher level education, and to give the skill provision to students to make self-development which is suitable for their ability and interest, as well as their environment condition. The ability in counting and measuring as well as reading and writing skill in elementary school are the basic for other skills development which is higher.

For instance, National Council of Teachers of Mathematics [1] propose that the development of mathematics skill of students in class K-2 usually can explore the similarity and the difference on two dimensions shape, in class 3 to 5 they are able to identify the characteristics of various rectangular, in class 6 to 8 they are able to check and make a generalization about the characters of certain rectangular, and in class 9 to 12 they are able to develop the logic argument to confirm the notion about certain polygon. The case gives the description that the mathematics skill develops from early age to mature or the mathematics skill develops from kindergarten level to high school level.

The result of Trends in Mathematics and Science Study [2] which is followed by Indonesian students of class VIII in 2011 shows that for mathematics field, Indonesia is in 38th place with the number of scores is 386 of 42 countries. This score decreases to eleven points from the rating in 2007. Program for International Student Assessment (PISA), hold the survey about students' skill and educational system, where the students' skills assessed in this survey are mathematics, reading and science skill which reflect the educational system in each country. The result shows that students' mathematics skill in Indonesia was ranked 64 of 65 countries or second

order from the bottom with the number of scores is 375.

The survey result indicates that Indonesian students' mathematics skill is still low both in content dimension and cognitive dimension. The content dimension assessment on the domain: numbers, algebra, geometry, data and probability. Meanwhile, on the cognitive dimension assessment in domain: 1) knowledge, including the facts, concept and procedure which should be known by the students, 2) applying, which focus on students' skill in applying the knowledge and concept understanding to solve the problem or answering the question; 3) reasoning which focus on problem solving non-routine, complex context and taking many problem-solving steps.

According to the findings of a study conducted by the Mathematics Teacher Upgrading Development Centre Team in several Indonesian elementary schools, 51 percent of students have difficulty with counting aspects, 50 percent have difficulty with concept mastery, and 49 percent have difficulty with completing the story test. In 2002, based on the research result from Mathematics Teacher Upgrading Development Centre Team reveals that in several areas in Indonesia, most of elementary students have difficulty in solving the story tests and interpret the story test into mathematics model.

On the survey conducted, Indonesian team is only followed by students of class VIII or junior high school students, while students of class IV or elementary school students are not included in the survey. This case is necessary to be conducted a further research to assess the students' mathematics skill in class IV or in elementary school level. The assessment is crucial to be conducted in order to give the information about the development achievement of students' mathematics skill or how far students have mastered the competency that they want to reach as a result of the learning.

Until now days, there are still several problems in school related with the quality and assessment implementation activity, especially in elementary school level. This case relates with the goal, planning, implementation, result and follow-up of assessment result both it is conducted by teacher and school. Based on the research result by Djemari Mardapi, [3] reveals that there are still many teachers that are not guided on the test grid in making question. They tend to use the question taken from books. The research conducted by Kumaidi [6] reveals that teacher did not make the grid first in making test question. There are many teachers that do not utilize the data from assessment result to improve the learning process. The data is only used for giving a label for students that are pass or not pass.

In the theory of item response, its mathematical model has a meaning that the subject probability to answer the item rightly depends on the subjects' skill and the characteristics of the item. It means that the test participants with higher skill will have bigger probability to answer the question rightly rather than the participants test having low skill. There are three assumptions which underlie the theory of item response, such as unidimensional, local independence and parameter invariance [5-7].

The explanation of the three assumption: 1) unidimensional assumption means that every item of test only measures one skill. This assumption can be shown if only the test contains one dominant component which measure the subject's prestige. 2) local independence assumption will be fulfilled if the participant's answer toward the item test does not influence the participant's answer toward another item test. 3) parameter invariance assumption means that the characteristics of item test does not depend on the distribution parameter of participant's test skill and the parameter which is the character of the participant's test does not depend on the character of the item test [7].

On the theory of item response, the probabilistic approach was used to state the relationship between the participant's skill and the likelihood of correctly answering. In this theory, model of distribution used is logistical distribution, not normal distribution. This case is caused by the normal curve is bell-shaped [8], therefore the curve is not up monotone. This case causes the higher skill of the average has lower probability if it is compared with the probability of skill average. This case opposite with measuring principle, that the participant with higher skill has higher probability to answer rightly the instrument. On the calculation of area under the curve can be done by integration [9], due to it is the function the density of normal probability causes the integration becomes more complicated. This case causes the use of logistical model on the theory of item response.

There are three logistic models in the theory of item response, such as logistical model with one parameter, logistical model with two models and logistical model with three models. The differences of three models lay in the number of parameters used in describing the characteristic of item in model used. The parameters used are index difficulty, difference power index of item and pseudo guessing.

2. Research Method

This research is research and development research, namely the development of mathematics skill test in elementary school. The product that will be produced from this research is a mathematical skill test instrument in elementary school. The math skill test is used to identify the level of mathematics skill of elementary school students and measure the development of math skill of elementary school students. The product of the development of mathematics skill test for elementary school students is the instrument of mathematics skill test

of elementary school students in the form of multiple choice. Instrument development criteria are limited to knowledge domains based on the 2013 curriculum.

Instrument development model in the form of test using modifications of the Wilson Model and Antonio Model [10] with the following steps: (1) initial development of the test, (2) test trial and (3) broad-scale trial. Initial development consists of: design of tests and validation by expert's judgement. After the design of the test is complete, the test is validated by expert, if any items that are not yet eligible are revised first until the test is valid in content. The instrument was tested on students in grades III, IV, and V elementary school. Based on trial, unfit items were revised and fit items were assembled as fit mathematical test. This mathematical test is ready to be used for measuring, then continued the broad-scale trial process.

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 uses customized scales. Data has been already in the format and analysed by using BILOG-MG program. Obtaining the parameters of the test item based on the analysis of the test results, so that the necessary improvements to the details of the question can be made. Proof of validity based on internal structure can be verified by CTT and IRT [11]. 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) Match Item test (goodness of fit), c) Difficulty Level, d) Information Function and SEM, e) Item Characteristic Curve (ICC).

Reliability test in this stage used the formula of Cronbach-Alpha and analysed by using SPSS 22 program. Data of test results 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 scored dichotomous. Test of goodness of fit for the test in overall as well as each item using the BILOG-MG program. The second characteristic use the difficulty level or index difficulty by utilizing 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 $(-2{,}0 < b < 2{,}0)$.

3. Result and Discussion

One of the educational problems related to the quality of education is the low mastery of students toward the competence, as a result of students' inadequate assessment. The assessment system is not optimal because: (1) the quality of tests made by teachers is still inadequate, (2) the monitoring of the testing network in the area has not been conducted 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 tries to develop math skill test for elementary school students in grades IV and V. Development of mathematics skill test refers to Core Competency and Basic Competency based on curriculum 2013. The test is expected to be able to identify the level of math skill of elementary school students, measure the development of math skill of elementary school students, and compile a profile of the level of achievement of students' math skill.

Content Validity of the Instrument

Validity is classified into three types, such as: (1) validity of content, (2) validity of criteria (criterion-related) and (3) validity of constructs [12-15]. 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 [16]. 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 [12,14].

The validity of the content is determined by using expert agreement. The level of validity of content related to the field of study, also known as a measured domain, is determined by expert agreement in the field of study [17,18]. This case is caused by the measurement instruments, such as tests or questionnaires are proven valid only if the expert believes that the instrument is able to measure the mastery of the capabilities defined in the measured domain. Analysing the validity of the content uses the aiken formula.

The instrument can be stated valid if the expert believes that the instrument measure the things which will be measured. Expert judgement gives 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. Instrument Validated

No	Research Instrument	Number of Items
2	Instrument of grade III	40
3	Instrument of grade V	40

Table 1 is the instrument which will be validated. The instrument consists of instrument of grade IV and V. each instrument consists of 40 items of question. The result of validity instrument verification is described on the table

below.

Table 2. Calculation Index of Aiken's V

No	Instrument	Number of items	Index Average of Aiken V
2	Instrument of grade III	40	0.979
3	Instrument of grade V	40	0.988

Table 2 is an average table of calculation validity results by using Aiken's v Index in grade III and V test instruments, where each instrument consists of 40 problem items. Table 2 shows that grade III instruments have an average Aiken's v index 0.979, and grade V instruments have an average Aiken's v index of 0.988. The average of Aiken's v index in each instrument is quite high.

Reliability of Instrument

The reliability of a test is generally expressed numerically in a coefficient of $-1.00 \leq \rho \leq +1.00$ [19]. Mahrens & Lehman [20] stated 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 3. Instrument Reliability of Grade III, IV, & V

Instruments	Cronbach-alpha	Number of Items
Instrument of grade IV	0,883	40
Instrument of grade V	0,954	40

Table 3 is the estimation result of the reliability on the research instrument developed. This research estimates the reliability on the instrument in grade IV and V where on each instrument consists of 40 items of questions.

Table 3 explains that the number of reliability coefficient of instrument in grade IV is 0.951. The result of coefficient estimation of the reliability shows that three instruments are reliable to use in measuring the students' mathematics skill in elementary school grade IV and V.

The Result of Instrument Trial Test

The result of instrument trial test is initialled 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 stage or not. A precondition test in this research consists of unidimensional, local independency and invariant of parameter. This research is considered unidimensional, the assumption of local independence has been fulfilled and the assumption of parameter variance of students' skill can be concluded that it has been fulfilled.

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 value of x^2. The probability value of each item should fulfil $p>0.05$, otherwise revision is conducted 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 8. Analysis Result of fit model

Instrument	Criteria	Parameter of Logistics		
		PL 1	PL 2	PL 3
Instrument of grade IV	The item is fit	40	40	26
	The item is not fit	0	0	14
Instrument of grade V	The item is fit	40	38	16
	The item is not fit	0	2	24

Based on the results of fit model analysis in Table 8, the most suitable model for instruments is obtained in 1-PL model. 1-PL model is suitable for test instruments that have been developed due to the number of test items is more suitable for the 1-PL model than the 1-PL, 2-PL and 3-PL models. The suitable model used in the analysis is 1-PL, so the parameter that should be paid attention is the difficulty level (b) for each item. Items that do not fulfil the criteria of a good item based on the criteria of both parameters; it will be removed from the final product.

Estimation of Item Parameter

The analysis used to figure out the characteristics of a good item is by using 1 PL model. Items that fit the model with 1 PL are then reanalysed to figure out the characteristics of the item. The criteria for a good item according to model 1 PL are based on the difficulty level of the item (bi). The index of difficulty level ranges from -2 to +2 [7] (Hambleton & Swaminathan, 1985: 107). The estimation of item parameter in this research use MG Bilog program. The following is the result of estimated analysis of item parameters that have been done.

Table 9. Estimation of Item Parameter of Grade IV

Item	Parameter of difficulty Level (*b*)		Explanation	
	Grade 4	Grade 5	Grade 4	Grade 5
1	-0.977	-1.267	Accepted	Accepted
2	-0.102	-0.356	Accepted	Accepted
3	-0.621	-0.987	Accepted	Accepted
4	-1.298	-0.695	Accepted	Accepted
5	-0.428	-1.563	Accepted	Accepted
6	-1.594	-0.470	Accepted	Accepted
7	-0.225	-1.191	Accepted	Accepted
8	-0.788	-0.372	Accepted	Accepted
9	-0.054	-0.504	Accepted	Accepted
10	-0.508	-1.670	Accepted	Accepted
11	-0.274	-0.768	Accepted	Accepted
12	-1.298	-0.884	Accepted	Accepted
13	-0.740	-1.029	Accepted	Accepted
14	-1.434	-1.215	Accepted	Accepted
15	-0.664	-0.924	Accepted	Accepted
16	-0.665	-0.588	Accepted	Accepted
17	-0.740	-0.844	Accepted	Accepted
18	-0.078	-0.825	Accepted	Accepted
19	-1.631	-0.071	Accepted	Accepted
20	-1.464	-0.520	Accepted	Accepted
21	-1.034	-0.024	Accepted	Accepted
22	-1.668	-0.404	Accepted	Accepted
23	-1.094	0.038	Accepted	Accepted
24	-0.102	-0.024	Accepted	Accepted
25	-1.594	-1.597	Accepted	Accepted
26	-1.273	-0.421	Accepted	Accepted
27	-1.180	-0.606	Accepted	Accepted
28	-1.378	-0.404	Accepted	Accepted
29	-0.680	-0.009	Accepted	Accepted
30	-1.350	-0.623	Accepted	Accepted
31	-1.249	-1.216	Accepted	Accepted
32	-0.090	-0.787	Accepted	Accepted
33	-0.495	-1.096	Accepted	Accepted
34	-0.756	-1.597	Accepted	Accepted
35	-1.464	-1.293	Accepted	Accepted
36	-0.415	-1.216	Accepted	Accepted
37	-1.434	-0.806	Accepted	Accepted
38	-1.249	-0.945	Accepted	Accepted
39	-0.299	-0.071	Accepted	Accepted
40	-0.788	-0.308	Accepted	Accepted

Based on the result of item parameter in the test instrument of grade IV and V show that the overall items are on good category which are between -2 to 2. This case indicates that the overall items are accepted and reliable to measure the students' mathematics skill of elementary school.

4. Conclusions

Based on the results of the development and discussion, the development of elementary school math skill test, the average of Aiken value for the instrument of Grade 4 is 0.979. Based on the data, it can be stated that all items proved valid reviewed from the validity of the content. The Aiken value range for the instrument of grade 5 is 0.988. Based on the data, it can be stated that all items proved valid reviewed from the validity of the content. Based on the results of the analysis showed that the coefficient of reliability instrument of grade 4 is 0.883, package instrument of grade 5 is 0.954, it is concluded that all instruments developed can be stated reliable.

Precondition test result showed that Unidimensional test is fulfilled due to the test is proven to measure only one dominant dimension of the same ability. Local independence assumption tests are also fulfilling due to the covariance value among the interval of skills are small or close to zero. The result of calculation of correlation between the difficulty level of the response includes in high category, so that the assumption of invariance of capability parameters is fulfilled. Based on the analysis of two instrument of fit model results, it is suitable for 1PL model only, so in the parameter estimation of the overall items package only estimate on the model 1PL or parameter b (difficulty level) only. Based on the results of the analysis of test item parameters of grade 4 and grade 5, it shows that the overall items are in good category that are between -2 to 2. It shows that all items are accepted and reliable to be used to measure the development of math skill of elementary school students.

REFERENCES

[1] NCTM. (2000). Principles and standards for school mathematics. Reston, VA: NCTM, Inc.

[2] OECD. 2013. PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy. OECD Publishing.

[3] Djemari Mardapi. (1999). Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional. Pidato Pengukuhan Guru Besar. Yogyakarta: Universitas Negeri Yogyakarta.

[4] Kumaidi. (2004). *Sistem asesmen untuk menunjang kualitas pembelajaran*. Jurnal pembelajaran, 27, 93-106.

[5] Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.

[6] Hambleton, R.K., Swaminathan, H., & Rogers, HJ. (1991). Fundamental of item response theory. Newbury Park, CA: Sage Publication Inc.

[7] Retnawati, H. (2014). Membuktikan Validitas Instrumen dalam Pengukuran. Diambil dari: http://evaluation-edu.com /wp-content/uploads/2014/10/2-Validitas-heri-Retnawati-uny.pdf

[8] Swasti Maharani, Toto Nusantara, Abdur Rahman As'ari, Abdul Qohar (2019). Analyticity and Systematicity Students of Mathematics Education on Solving Non-routine Problems. Mathematics and Statistics, 7(2), 50 - 55. DOI: 10.13189/ms.2019.070204.

[9] Alec John Villamar, Marionne Gayagoy, Flerida Matalang, Karen Joy Catacutan (2020). Usefulness of Mathematics Subjects in the Accounting Courses in Baccalaureate Education. Mathematics and Statistics, 8(1), 27 - 31. DOI: 10.13189/ms.2020.080103.

[10] Suprapto, E., Sumiharsono, R., & Ramadhan, S. (2020). The Analysis of Instrument Quality to Measure the Students' Higher Order Thinking Skill in Physics Learning. *Journal of Turkish Science Education*, *17*(4), 520-527.

[11] Vendramini, C.M.M. & Silvia, M.C.R. (2011). Application of item response theory in the attitudes evaluation. Diambil tanggal 4 April 2015, dari http://www.tsg.icme11.org/document/get/492.

[12] Nunnally, J. (1978). Psychometric theory (2nd ed.). New York: McGraw Hill.

[13] Allen, M. J. & Yen, W. M. (1979). Introduction to measurement theory. Belmont, CA: Wadsworth, MC.

[14] Fernandes, H. J. X. (1984). Evaluation of educational program. Jakarta: National Education Planning, Evaluating and Curriculum Development.

[15] Woolfolk, A. E. & McCune, L. N. (1984). Educational psychology for teachers. Englewood Cliffs, NJ: Prentice Hall, In.

[16] Lissitz, W. & Samuelsen, K. (2007). *Further clarification regarding validity and eduction. Educational Researcher*, Vol. 36, No. 8, pp. 482-484.

[17] 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.

[18] 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.

[19] Retnawati, H. (2016). Validitas, Reliabilitas, & Karakteritik Butir. Yogyakarta: Parama Publishing

[20] Mehrens, W.A. & Lehmann, I.J. (1973). Measurement and evaluation in education and psychology. New York: Hold, Rinehart and Wiston, Inc.