International Journal of Instruction
e-ISSN: 1308-1470 ● www.e-iji.net

July 2018 ● Vol.11, No.3
p-ISSN: 1694-609X
pp. 375-390

Received: 06/11/2017
Revision: 24/03/2018
Accepted: 28/03/2018

Developing a Testlet Model for Mathematics at Elementary Level

Syukrul Hamdi
Dr., Hamzanwadi University, Indonesia, *syukrulhamdi@hamzanwadi.ac.id*

Badrun Kartowagiran
Prof. Yogyakarta State University, Indonesia, *kartowagiran@uny.ac.id*

Haryanto
Dr., Yogyakarta State University, Indonesia, *haryanto@uny.ac.id*

The purpose of this study was to develop a Mathematics test instrument testlet model for a classroom assessment at elementary school. Testlet Model is a group of multiple choice question acquiring similar information with different grade of responses model. This research was conducted in East Lombok, Indonesia. The design used was research development model. The study involved 17 doctoral students as peer reviewers, 19 teachers as testlet preparation workshop participants and 711 students for testing the testlet mathematical test instrument. Data collection techniques used were documentation techniques, interviews and questionnaires. Expert validation results were analyzed using Content Validity Index (CVI). Analysis of the trial results was using items response theory with a graded response model of two logistic parameters. Independent sample t test was used for comparison of tests. The research developed a Mathematics test instrument testlet model successfully for a classroom assessment consisting of the testlet model test design, grading guide, test instrument preparation guide and mathematics test instrument testlet model or an accurate and reliable classroom assessment. The result of the analysis shows that there was different result between assessments using testlet compared to multiple choices scoring, with t value equal to 7.864.

Keywords: testlet model, mathematics, classroom assessment, elementary school, test

INTRODUCTION

Measurement is a necessary process to improve the quality of education. Positive change of a quality is always shown by a quality of output which can also be seen through the measurement. Measurement in the field will be related with several things happening during the learning process, especially evaluation and assessment. Griffin and Nix (1991) describe that a measurement, assessment and evaluation have hierarchal

Citation: Hamdi, S., Kartowagiran, B., & Haryanto (2018). Developing a Testlet Model for Mathematics at Elementary Level. *International Journal of Instruction*, 11(3), 375-390. https://doi.org/10.12973/iji.2018.11326a

characteristics. Measurement is comparing observation result with criteria while assessment is explaining and interpreting the measurement result. The hierarchal characteristics of these processes show that every evaluation activity involves a measurement and assessment (Mardapi, 2008).

The fact in the field shows that the ability of most teachers in preparing an assessment instrument is still low. It's supported by a study on *professor go to schools* activity performed by distributing questionnaires to teachers. The result showed that 79 percent of teachers had difficulty making an assessment instrument (Rohmawati, 2013). The result is supported by a study by Retnowati, at al., (2016) which states that "*the knowledge assessment became the single core competency and the teachers had few difficulties in implementing the assessment"*. Similarly, Kartowagiran and Jaedun (2016) also state that based on research only a small number of teachers regularly make assessment, and only a few teachers prepare assessment instrument. It's because most teachers feel they lack training so they don't fully understand the materials, especially assessment material.

One of the assessment models applied is the classroom assessment. The classroom assessment is often applied because it's consistent with the curriculum implementing in Indonesia today, particularly those related with several aspects which are the main objectives of education. Classroom assessment according to Russel and Airasian (2012) is a process aimed to collect, synthesize, and interpret information in making a decision in the classroom. Tierney (2006) states that the usage of a classroom assessment is aimed to promote the greater students' learning, especially from the latest research. Seven principles of effectiveness in the classroom assessment include requirement for classrooms. Therefore, the classroom assessment on quality of education can be performed by various assessment instruments. The assessment instruments could be test or non-test. Test instrument is categorized into objective test and non-objective test.

One of the regions which use both test types is East Lombok Regency. East Lombok is one of the centers of the measurement in various fields because it has the highest population compared with other regions in West Nusa Tenggara Province-Indonesia. Based on preliminary study in East Lombok Regency, some multiple choice tests used by teachers in semester exam can't reveal students' actual ability. A study on the implementation of curriculum 2013 (Abrory and Kartowagiran, 2014) concludes that the quality of the assessment of math lesson is poor. One of the factors of the problem lack of item analysis theoretically and empirically. The opinion was supported by the result of the analysis of items made by teachers with 6261 students of elementary school as respondents from A, B, and C accredited schools in 20 sub-districts in East Lombok. It shows that from 40 items used for math semester exam 2015 only 4 items or 10% are satisfactory, 42.5% unsatisfactory and 47.5% very unsatisfactory (Dikpora Kabupaten Lombok Timur, 2015). The analysis result means that the test instrument used was not fulfilled a good test standar.

Mathematics is a subject covered in both junior high school and elementary school national examination. According to Humenberger (1997), "students should see and experience that math is a language which can translate many problems, and help us solve and resolve problems in various cases". It means that student's math ability can be used

to solve various problems and difficulties they face in learning various sciences, especially natural sciences. De Lange's (Van den Heuvel-Panhuizen, 1996) classified mathematics education objective into three levels, i.e. low level, medium level, and high level objectives. De' Lange's also presents levels of understanding as a pyramid shown. The low level objective of mathematics education is to make student master mathematics concepts related to object knowledge, definitions, technical skill, and standard algorithm, such as addition of integers, fractional numbers, etc.

At the second level which is the medium level, mathematics education objective is characterized by students' understanding on two or more concepts, making relation, integration, and problem solving are common terms at this level. At the third level which is high level, students' understanding is characterized by abilities to work with complex materials such as mathematical thinking and reasoning, communicative, critical attitude, creative, interpretative, reflective, generalizing, and mathematical. Haylock and Thangata (2007) state that problem solving can be defined as the ability of students to use their mathematics thinking and knowledge to overcome the problems presented and to achieve the result. So, problem solving requires creativity to think scientifically and to use logical reasoning.

Learning outcome assessment by teacher is performed continuously to observe process, progress, and improvement of outcome using test instrument or non-test instrument. One of the instruments is test related with a reality or realistic problem. In reality, some teachers don't fully understand the procedure or application of the test preparation in the field. Therefore, review of various references and theories by measurement experts is necessary to be basis for educators in the instrument development and outcome analysis.

There are two important instrument analyses which are qualitative analysis and quantitative analysis. Quantitative analysis uses the classical test theory and modern test theory which is known as Item Response Theory (IRT). It's found in the field that teachers and academicians mostly use classical test theory and multiple choice dichotomous scoring. One of the weaknesses of the classical test theory approach is lack of information on the response of each examinee on each item (Mardapi, 2004). The alternative approach is the item response theory approach for polytomous scorings.

Polytomous scoring usually uses essay test and testlet model test. Testlet is a group of multiple choice items which reveal the same information (Weiner and Kiely, 1987). Furthermore, according to Wainer at al., (2007) testlet is a type commonly used to solve problems, especially for efficient time in giving an individual task or exercise to students to determine various types of stimulus. The basic idea is processing stimuli from examinee, who must solve several items which reveal the same information. The items enable measuring the same achievement outside of the measured property by using the test as a whole. This is supported by Paek at al., (2009), who state that in the past few years, the model has evolved to accommodate test with testlet structure in which a series of item has simultaneous stimulus. Similarly, Yue and Hong-Yun (2012) say that testlet consists of a group of multiple choice items based on simultaneous stimulus. Based on expert opinions above, it's concluded that testlet must be multiple choices and multiple choice test isn't necessarily testlet because testlet consists of at least two

multiple choice items based on simultaneous stimulus. It differentiates multiple choice test and testlet test.

Edward (2010) reveals that the testlet model design is used in education and research to see the response of a test taker for a question model where the answer is from one paragraph. Therefore, the testlet model test is very useful to show the comparison of each response of test takers. In other words, the testlet model is quite simple but the provided information is good enough. Based on the theoretical and empirical review by Susongko (2010), in terms of scorings, the testlet is more practical than description because the scoring can be more objective and polytomous.

Expert definitions and opinion based on previous studies on testlet indicate that testlet can be a helpful alternative in measurement. The role of testlet will meet expectation if developed based on valid theories from experts and supported by previous studies which can be used as references in making comparison for better development. The fact and problem in the field, particularly in the research location, are that testlet has never been studied specifically. Testlet model test is very rarely or never used by math teachers in classroom assessment. The common assessment/test instrument is essay test and multiple choice test with dichotomous scoring.

An essay test has many advantages, but it takes longer time for scoring. Moreover, in general a teacher handles 32 students if not more per classroom, so the possibility of making errors in assessment is quite big. Braun at al., (Myford and Wolfe, 2009) argues that assessor's weakness is systematic changes often happen in assessing from time to time. Moreover, the weakness of essay test is tidiness of handwriting. It happens when assessor is distracted by tidiness of handwriting, making answer looks good, but when they read carefully its content, they didn't answer the question (Mardapi, 2004). The advantages and weaknesses of objective tests are nearly the same. Although multiple choice model has many advantages, it only measures student's score and cannot be used to identify student's weakness in lesson.

The condition in the field is a problem which should be proven empirically considering the testlet model test is a rare assessment model in a classroom assessment. The specification of the testlet can provide broad view on student's learning development. Various facts and existing theories are a powerful reason for the importance of studying testlet. The complexity of meaning which can be obtained from using testlet in classroom will provide significant benefit, especially if the developed items or test instruments are related with real world problem.

METHOD

Development Model

This research used a development research referring to the test development model from Mardapi. According to Mardapi (2008) there are nine steps that are needed in developing the test result for learning: (1) preparing the test specs, (2) writing test questions, (3) studying the test, (4) assessing the test instrument (5) analyzing the items, (6) improving the test, (7) assembling the test, (8) carrying out the extended test / trial, (9) analyzing and interpreting the test results.

The arrangement of test specifications based on test objectives for classroom assessment goes to the formative assessment category, the formulation refers to the curriculum applied in Indonesia two of which are 2013 curriculum and KTSP curriculum. The developed instrument is an instrument that combines the strengths contained in multiple choice tests and test descriptions called testlets. The tests are developed using realistic problems so that they are easily understood by the students. Specification of development result test was in the form of lattice. In the second stage was to proceed by writing a problem based on the grid that has been generated from the development process. Problems that have been made then reviewed by three experts of measurement and assessment of education namely Prof. Djemari Mardapi, Ph.D, Prof. Dr. Badrun Kartowagiran, Dr. Haryanto and four experts of Mathematics education; Dr. Anak Agung P, Dr. Dhoriva UW, Dr. R. Roesnawati, M.Si, and Dr. Zamsir. The study also involved 17 doctoral students as peer reviewers. Test instruments that have been in the study conducted limited trials and analyzed to determine the quality of the test to get a valid and reliable test. The sixth stage froze the revision based on the analysis that was done then assembled the test. Test instruments assembled in accordance with the purpose of further tests are used for extended trials to be analyzed and interpreted. The end result obtained in the form of a high quality testlet mathematical test instrument and its scoring guidelines. The results of the development are disseminated to the users of the elementary school teachers through workshop preparation of mathematical testlet test models. From the workshop activities Mr / Ms teachers were asked to assess related to the teacher's assessment and the use of Mathematics test instrument testlet model.

Participants and Research Sample

This study involved 7 experts for studying the mathematical test models. The experts are consisting of three measurement experts and four mathematics education experts drawn from three universities representing three provinces in Indonesia. This study also involves 17 doctoral students (concentration of evaluation, measurement and concentration of mathematics education) as peer reviewers who made in one forum that is Mathematics test instrument testlet model. The study was conducted for examining the test-model, test instruments and scoring models as well as the final assessment used in the developed model. The study also involved 19 primary school teachers (R1, R2, R3,., R19) as workshop participants and interviewees related to student learning outcomes and given teacher assessment sheets related to Mathematics test instrument testlet model.

Curriculum implementation in the Indonesian education focuses on two types of curriculums, namely KTSP Curriculum and the 2013 Curriculum. The materials included in KTSP Curriculum remain specific and not integrated with the other subjects, while in the 2013 Curriculum, the nature of the materials is thematic-integrated. The test phase for this Mathematics test instrument testlet model involves schools that are implementing the 2013 curriculum and the KTSP curriculum. The population of this study was the all elementary school students in East Lombok Regency, NTB Province of Indonesia. The sampling technique used was Classified Random Sampling, the sample was taken randomly based on geographic location such as district school, border / town of sub-district and village / mountains. In a limited trial process involving 159 students

from four elementary schools as a sample of two curricula applied in Indonesia. The sample used for the extended trial of 552 drawn from 14 primary schools consisting of 273 students from elementary school implemented the KTSP curriculum (1 district school, 3 border schools, 3 rural schools) and 279 of the primary schools implementing the 2013 curriculum (2 district city schools, 3 border schools, and 2 rural schools). The total sample used from the limited to expanded trial is 711 students (352 from elementary school based curriculum KTSP and 359 from elementary school with 2013 Curriculum). The sample used can represent all students in the East Lombok Regency of West Nusa Tenggara Province in Indonesia.

Table 1
Test Participants of Elementary School

	2013 Curriculum		School Based Curriculum (KTSP)	
No	School	Number of Students	School	Number of Students
Limited Trial Participants				
1	SDN 4 Pancor	43	SDN 2 Pancor	30
2	SDN 7 Danger	37	SDN 2 Sandubaya	49
Extended Trial Participants				
3	SDN 3 Selong	74	SDN 1 Sakra	49
4	SDN 3 Pancor	52	SDN 2 Rakam	23
5	SDN 3 Rempung	35	SDN 2 Aikmel	64
6	SDN 1 Anjani	29	SDN 3 Sikur	43
7	SDN 3 Masbagik Timur	29	SDN 4 Lendang Nangka	40
8	SDN 3 Aikmel	19	SDN 1 Pringgasela	35
9	SDN 7 Lendang Nangka	41	SDN 5 Suralaga	19
Total		359	Total	352

Data Collection and Analysis

Data collection techniques used were documentation techniques, interviews, tests and questionnaires. Expert / expert assessment results were analyzed by using Content Validity Ratio (CVR) and Content Validity Index (CVI). It was used to determine the conformity of the items towards indicators and the conformity of indicators towards the basic competencies within the curriculum. The results of CVR and CVI were estimated as evidence that the developed instrument meet the validity of its content. Valid instruments were tested for legibility and limitation to know the functionality of distractors, differentiation and difficulty using ITEMAN program. The revised instrument was used for the expanded trial, the test results were analyzed using grain response theory with a tiered response model of two logistic parameters (2-PL GRM). The participant's response to the first item with the GRM model was scored as category, k = 0,1,2,. . ., m is the number of steps in correct completion of item j, and an index of difficulty in each sequential step. With GRM analysis, the researcher could see the participant's response pattern and knew the position of the participant's weakness in the learning through assessment as the basic for learning improvement.

The approach used in the grain response theory was Marginal Maximum Likelihood (MML), and to know the parameter of the item using Bock and Liebermen method. The value of the test information function and the standard error of estimation was based on

the estimation of the grain parameters. The value of the test device information function (Hambleton, Swaminathan, and Rongers, 1991) was calculated by the following formula.

$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$ The value of the information function was popular in the grain response theory, the grain information function is a method to explain the strength of a grain on a test device. The function of the grain information states the strength or contribution of grains to the test in uncovering the latent trait as measured by the test. With the grain information function, it provided information about which items are matched with the model so it might help in the items selection. The standard of error estimation can be determined from the value of the information function. The standard error of estimation was calculated by the following formula.

$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$ with I_i (θ) = the information function of the first and SE items (θ) denotes that the raw error of estimation and I (θ) function of the test information. SE (θ) in IRT analog with the standard of error measurement (SEM) goes to classical theory. The calculations were performed in a skill level scale of -4.00 + 4.00 with 0.1 intervals. Computing was conducted by using Microsoft Excel program. To know the effectiveness of the instrument was proved by comparing the scoring results using the test model test instrument with a multiple choice scoring model dichotomous using t-test (independent sample t-test). Results of teacher and student responses towards the guidebooks and instruments and observations during the research process were analyzed using qualitative descriptive.

FINDINGS

The developed model of the Mathematics testlet test is a testlet consisting of three arrangements with pholytomus scale, scale 0-3 (0,1,2, and 3) or four categories. The grille for classroom assessment should contain Basic Competency, Indicator and Question Item Number. The example of grille Mathematics test instrument that will be developed is as followed.

Table 2
The Grille for Mathematic Test Istrument Testlet Model

Basic Competency	Indicator	Testlet Item Number	Testlet Sub-Item Number	Question Number
		1	1.I	1
			1.II	2
			1.III	3

The design used in this development was consistent with three levels of objectives of mathematics education, i.e. low level, medium level and high level, which are interpreted as easy, medium, and high questions of each competence tested in 1 testlet item group. Making essay mathematic questions based on the plan aims to help teacher because, according to teachers, essay question is the easiest to make. The testlet designed in this research used three questions, so the essay questions were adjusted with the provisions made in the pan, so the process of answering went through three steps or more. Example Question: What is the area of the front side of the Aceh traditional house

made by the teacher which consists of two two-dimensional figures (rectangle and triangle)?

The next step is to answer the essay questions and identifying steps or procedure to get the final answers of the questions. For example in Example A: To get overall area, the length of each two-dimensional figure should be determined if the area of a two-dimensional figure is known of the areas of the two-dimensional figures should be determined separately if all lengths are known. Making questions based on the steps or procedure. To gradually make testlet questions from easy to difficult, examine which part is easier in each step to get answers from essay questions. In this case, determining the area of a rectangle is easier than determining the area of a triangle because all sizes in the figure are clear for the rectangle. If the area of the rectangle is known, the length of the rectangle also can be determined to answer the next question. After making questions from the procedure to get answers, make the questions into a testlet which consists of three questions, which are determining the one of side length of the rectangle based on area which is known, determining the area of the triangle, and determining the combined area of the two-dimensional figures. To be clear, observe the following example.

Exemple. Answer questions 1-3 based on the following illustration.

One of Indonesia's riches is Aceh traditional house, as shown in Figure 1a The teacher has drawn the front side of the traditional house which consists of a combination of two two-dimensional figures as in Figure 1b. Before drawing door and window, the teacher asks the students to calculate the front side of the Aceh traditional house based on the sizes given by the teacher in Figure 1b.

Figure 1a

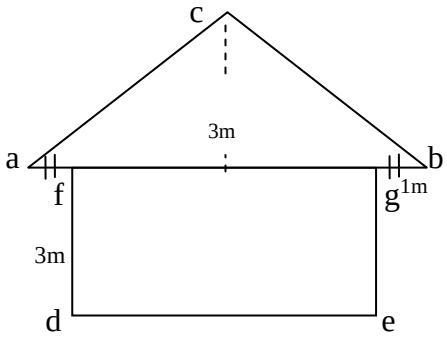


Figure 1b

Figure 1
Aceh Traditional House

1. If the area of the rectangle in the front side of Aceh traditional house is 18 m^2 the length of $\overline{de}$ is . . . .
   A. 3 m^2 B. 6 m^2 C. 18 m^2 D. 24 m^2
2. The area of the triangle in the front side of Aceh traditional house made by the teacher is. . . .
   A. 6 m^2 B. 8 m^2 C. 12 m^2 D. 24 m^2

3. The area of the front side of side of Aceh traditional house made by the teacher which consists of the combination of two two-dimensional figures (rectangle and triangle) is . . . .
   A. 12 m^2 B. 22 m^2 C. 24 m^2 D. 30 m^2

The test instrument developed using level form of easy question category to the difficult question category. If we refer to the question elaboration, the multiple-choice arranging the testlet is part of some phases that should be finished in answering the question appropriately. The student who can give a right answer in item 2 is supposed to answer the question in number 1, also for the case in item 3. Scoring rubric used in testlet is presented in the following Table 3.

Table 3
Assessment Rubric of Mathematics Test Instrument Testlet Model for Classroom Assessment in Elementary School

No	Assessment Criteria	Testlet Item			Score	Description
		I	II	III		
1	The students are not able to finish the first question, second, third in the same group of testlet item appropriately	0	0	0	0	Students do not comprehend and understand the material well.
2	The students are not able to finish the first question, and the third but able to finish the second question in the group appropriately	0	1	0	0	Students tend to guess
3	The students are not able to finish the first question, but they can finish the second and third question appropriately in the group item	0	0	1	0	Students get the answers by guessing
4	The students could not finish the first question, but they can finish the second and the third appropriately in the group of the testlet item	0	1	1	0	Students do not do the test carefully so the loss the material
5	The students could not finish the second and third question but they could finish the first question appropriately within the group of testlet item	1	0	0	1	Students do not understand the basic concept of question
6	The students could not finish the second question but they can finish the first and third appropriately within the group of the testlet item	1	0	1	1	Students understand the concept but do not do the test carefully
7	The students are not able to finish the third question but they could finish the first and second question appropriately within the group of testlet item	1	1	0	2	Students understand the main concept, the middle level of test given
8	The students could finish the first, second, and the third appropriately within the group of testlet item	1	1	1	3	The Student comprehend and understand the task material well

Description:

I = the easy question from the group of testlet item, II = the middle question from the group of testlet item, III = the difficult question from the group of testlet item

Assessment on design, preparation guide and rubric showed that four experts or 57.14% stated that they were very appropriate/good, two experts or 28.57% stated that they're appropriate/good, and one expert or 14.28% stated that they're adequately appropriate/good. Based on expert opinions, it's concluded that the assessment design, preparation guide and rubric were appropriate. The estimation content validity show that the BC accordance CVI (Content Validity Index) value and indicator is 0.91 and 0.83, which means that the two-instrument grille developed, is valid. The value of CVI accordance item towards indicator achieved 0.96 and 0.85 which means that the two instrument package of mathematic testlet model developed is valid based on the experts.

In general, the characteristic of Mathematics test instrument testlet model for the 2013 curriculum which has been developed with -0.642 to 2.668; different power from 0.432 to 0.941 and guessing has 0 value since the model selected 2PL GRM which means that the guessing value is 0. The characteristic of mathematic testlet test model for 2016 curriculum which has been developed has -0.441 to 3.603 level of difficulties; defirrence power 0.312 to 0.808 with the guessing value 0 due to model selected 2PL GRM which means that the guessing value is 0. The selection of 2PL GRM in the research based on the conformity model test result. In general the ability of Elementary School students goes to middle category which exist between -2 to +2. The scoring guide used is fix based on the assessment of experts and book guide of Mathematics test instrument testlet model because it completed with some ways and example directly.

Scoring Guide Mathematics Test Instrument Testlet Model

1. Open the scanned testlet file using Excel Microsoft program, then the MENU picture will appear as below.

Figur 2
Menu Picture

2. In the menu field are available five main options consisting of:
   - STUDENT ANSWER (JAWABAN SISWA)
   - EARLY SCORES (SKOR AWAL)
   - TESTLET SCORES (SKOR TESTLET)
   - DESCRIPTION (DESKRIPSI)
   - GRAPHS (GRAFIK)

In addition, two additional menu options are available in the left position that read USER INSTRUCTIONS and an additional option at the bottom of the column that says EDIT CODE DESCRIPTION.

3. To use the testlet scoring guidelines, select or click the USER INSTRUCTIONS menu.
4. After that, will appear student answers column. In this column is available FILL DATA menu with display as follows.

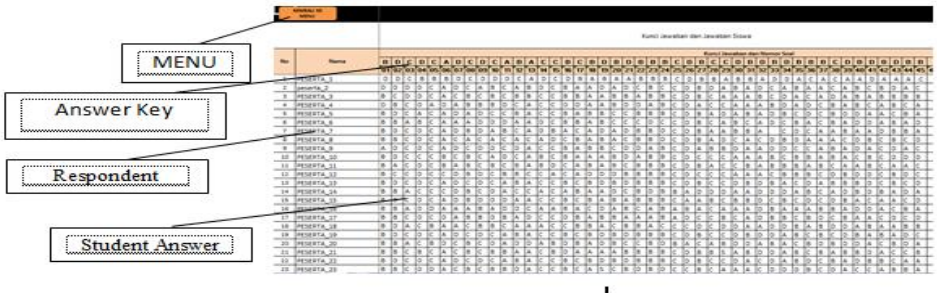
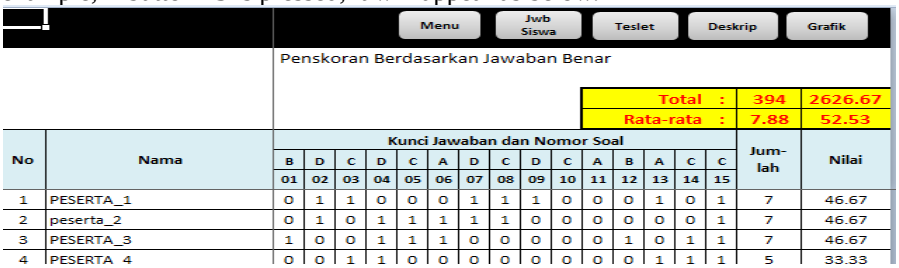


Figure 3
Fill Data Table

5. Next, enter the answer key in the provided field (Answer Column and Problem Number) according to the number of questions provided.
6. Next, write the name of the examinee (student) in the Name Column.
7. Then, enter the student's answer in the assigned column according to the number of student answer and answer keys.
8. After all student answers are entered, press MENU at the top right to return to the start screen.
9. Early Scores in this guide are scores obtained by students, according to the number of correct answers or dichotomous scores of all questions. At the beginning SCORE there are four options namely 

. The choice serves to facilitate the teacher and to accommodate the number of desired questions. For example, there are 5 grains of testlets consisting of 3 items of constituents, so the teacher can use 15 or simply by pressing or clicking the button 15. The same is true for buttons 30, 45 and 60. Everything is tailored to the needs of the teacher. For example, if button 15 is pressed, it will appear as below.

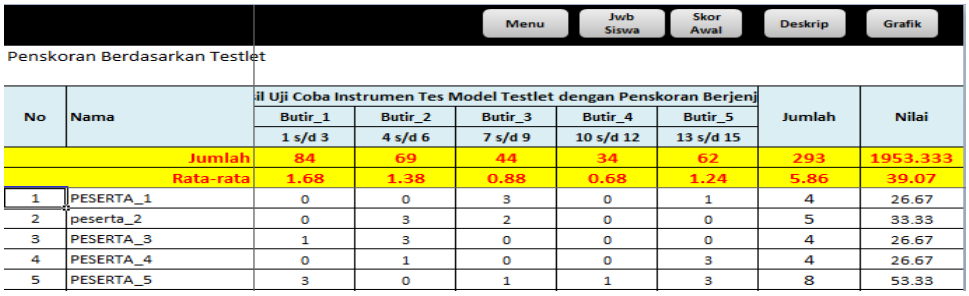


No	Nama	Kunci Jawaban dan Nomor Soal															Jum-lah	Nilai
		B	D	C	D	C	A	D	C	D	C	A	B	A	C	C		
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15		
1	PESERTA_1	0	1	1	0	0	0	1	1	1	0	0	0	1	0	1	7	46.67
2	peserta_2	0	1	0	1	1	1	1	1	0	0	0	0	0	0	1	7	46.67
3	PESERTA_3	1	0	0	1	1	1	0	0	0	0	0	1	0	1	1	7	46.67
4	PESERTA_4	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	5	33.33

Figure 4
Early Scores

To return to the beginning just press the MENU button. So do with other menus. For example, after viewing the score of students, teachers want to directly see the graph of the score then it can be directly viewed by selecting or pressing the graphic menu.

10. The TESTLET SCORE that appears on the app display is a score based on the validated scores of testlet scores. There are four options with the same conditions as point 9 instructions. For example, when button 15 is pressed, the display appears as below.



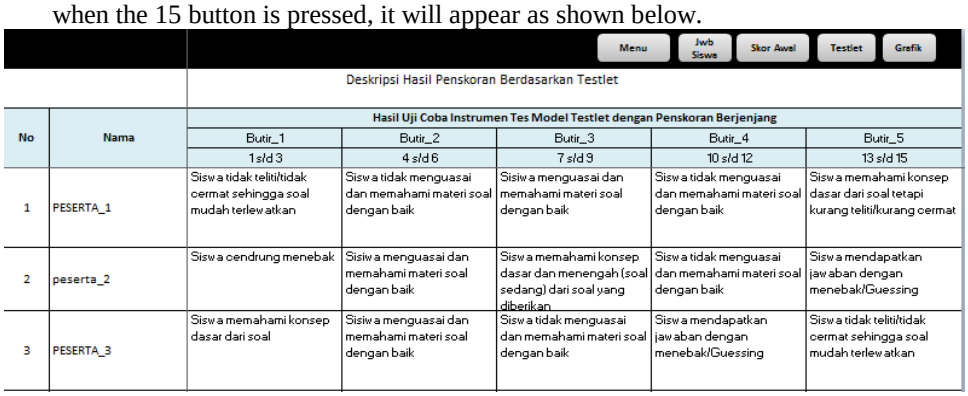
Menu | Jwb Siswa | Skor Awal | Deskrip | Grafik

Penskoran Berdasarkan Testlet

No	Nama	il Uji Coba Instrumen Tes Model Testlet dengan Penskoran Berjenj					Jumlah	Nilai
		Butir_1	Butir_2	Butir_3	Butir_4	Butir_5		
		1 s/d 3	4 s/d 6	7 s/d 9	10 s/d 12	13 s/d 15		
	Jumlah	84	69	44	34	62	293	1953.333
	Rata-rata	1.68	1.38	0.88	0.68	1.24	5.86	39.07
1	PESERTA_1	0	0	3	0	1	4	26.67
2	peserta_2	0	3	2	0	0	5	33.33
3	PESERTA_3	1	3	0	0	0	4	26.67
4	PESERTA_4	0	1	0	0	3	4	26.67
5	PESERTA_5	3	0	1	1	3	8	53.33

Figure 5
Testlet Scores

11. DESCRIPTION is exposure of student answer analysis based on scoring testlet. There are four menu options, according to the instructions number 9. For example, when the 15 button is pressed, it will appear as shown below.



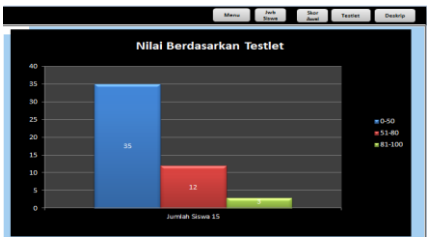
Menu | Jwb Siswa | Skor Awal | Testlet | Grafik

Deskripsi Hasil Penskoran Berdasarkan Testlet

No	Nama	Hasil Uji Coba Instrumen Tes Model Testlet dengan Penskoran Berjenjang				
		Butir_1	Butir_2	Butir_3	Butir_4	Butir_5
		1 s/d 3	4 s/d 6	7 s/d 9	10 s/d 12	13 s/d 15
1	PESERTA_1	Siswa tidak teliti/tidak cermat sehingga soal mudah terlewatkan	Siswa tidak menguasai dan memahami materi soal dengan baik	Siswa menguasai dan memahami materi soal dengan baik	Siswa tidak menguasai dan memahami materi soal dengan baik	Siswa memahami konsep dasar dari soal tetapi kurang teliti/kurang cermat
2	peserta_2	Siswa cendrung menebak	Siswa menguasai dan memahami materi soal dengan baik	Siswa memahami konsep dasar dan menengah (soal sedang) dari soal yang diberikan	Siswa tidak menguasai dan memahami materi soal dengan baik	Siswa mendapatkan jawaban dengan menebak/Guessing
3	PESERTA_3	Siswa memahami konsep dasar dari soal	Siswa menguasai dan memahami materi soal dengan baik	Siswa tidak menguasai dan memahami materi soal dengan baik	Siswa mendapatkan jawaban dengan menebak/Guessing	Siswa tidak teliti/tidak cermat sehingga soal mudah terlewatkan

Figure 6
Description Scoring Guide Mathematics Test Instrument Testlet Model

12. GRAPH consists of bar graphs and circle graphs. Bar graphs function to determine the number of students who scored under 50, between 51-80 and the score between 81-100 based on the scoring rubric of the testlet. On the menu of this graphic also four options are also available with the same provisions, based on the instructions number 9. For example, when the button 15 is pressed, then the display appears as follows.



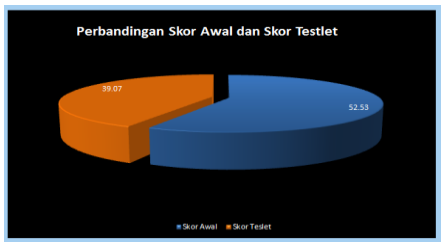


Figure 7
Graph Scoring Guide Mathematics Test Instrument Testlet Model

DISCUSSION AND CONCLUSION

The tool of Mathematics test instrument testlet model for the class assessment in elementary school consists of test instrument design, grille instrument, test instrument, scoring guide and test instrument arrangement guide. Instrument design of mathematic testlet model test which is developed is one item testlet consisting three item arrangement with pholytomus scale in 0-3 scale or four category. Testlet aims to combine the strengths of essay and multiple choice tests and complement the weaknesses of the tests. To get good testlet, the following steps can be taken: 1). Identification of objective and measurement area; 2). Mapping of Basic Competency (BC) and indicator; 3). Determination of limit of competency to be measured; 4). Determination of testlet design based on measurement to be used; 5). Making of grille of testlet model mathematic test instrument; 6). Making of essay mathematic questions based on the plan; 7). Answering the essay questions and identification of steps or procedure to get the final answers of the questions; 8). Making of questions based on steps or procedure using multiple choice questions, and 9). Compilation of the questions into a testlet after making questions from the procedure to get answers.

One test item called testlet consists of three items multiple choices so that why in determining each arrangement item should be well planned in the Mathematics test instrument testlet model. The arrangement of grille refers to the result of basic competence determination and the indicator of curriculum used. Each basic competence (BC) consists of one minimum indicator each of which should have 1 item testlet arrangement. The grille product of Mathematics test instrument testlet model for a classroom assessment in elementary school using 2013 curriculum has been revised based on the expert recommendation.

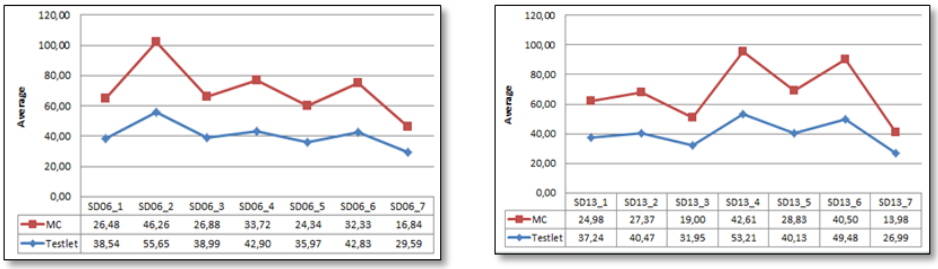
The scoring guide used is also good based on experts and test instrument guide in mathematic testlet model test according to the teachers are also good, simple and easy to understand since it has strategy and example at the same time. The quality of test instrument in mathematic testlet model test for classroom assessment in elementary school can be seen from: 1) experts assessment, 2) the characteristic of difficulties and differential power of test instrument testlet model and 3) the value of information function, wrong measurement and reliability of test instrument in the testlet model. The characteristic of the instrument tool is using the item response theory that has been assessed involving uni-dimension assumption, local independency assumption and

invariance parameter assumption. These three assumptions has been completed both for mathematic testlet model test for KTSP curriculum and test instrument of mathematic testlet model test using 2013 curriculum.

The important parameter value of test participants are between $-4 \leq \theta \leq +4$ and in the scale $-2 \leq \theta \leq +2$,ideally where the difference power parameter should be positive$(a_i > 0)$. According to Hambleton at al., (1991) the parameter of difficulty level in each items appears in $(-\infty \; sampai \; dengan + \infty)$ and will be good or ideal if $-2 \leq b_i \leq +2$.

The research result showed that all difference power goes to a good category even best category both for KTSP curriculum and 2013 curriculum. On the other hand, the difficulty level of question for 2016 curriculum question has four unideal question or goes to bad category such as presented in number 4,7,14 and 15. The fourth question are difficult and it can be answered by the student with high ability $(\theta > +2)$.

The next step after looking at the question characteristic is estimating the value of information function, miss measurement and the reliability of test instrument in mathematic testlet model test for class assessment in elementary school that are implying 2016 curriculum and 2013 curriculum. The analysis result show that the value of information function meet on the ability scale -1.8 and +3.5. Between this two ability, the item has information function value that is higher compared to error standard or its miss-measurement. On the other side, under -1.8 and up to +3.5 its miss-measurement is higher compared to its information function value or given information. Meanwhile, for the 2013 curriculum this function meet on the ability scale of -1.7 and +3.6. Between these two ability items, this item has a higher information function value than error standard and its miss-measurement. Moreover, under -1.7 and up to +3.6 its miss-measurement is higher than its information function value of the given information. Next, for the mathematic test instrument reability for KTSP curriculum is 0.822 and instrument test of mathematic testlet model test for 2013 curriculum is 0.808. This result show that the test instrument in mathematic testlet model test for class assessment in elementary school using KTSP and 2013 curriculum are reliable.



KTSP curriculum 2013 curriculum

Figure 8
Graphic difference between the average of testlet and multiple-choice base on the KTSP curriculum and 2013 curriculum.

Referring to the picture presented, it is clear that there is difference between the average value achieved using different scoring; testlet and multiple-choice. This difference describes that the higher difference between testlet and multiple-choice, the guessing level of student is higher.

The information that can be propose from this analysis result is the different result between assessment using mathematic testlet model test and assessment with multiple choice mathematic test instrument. The result of the analysis shows that there was different result between assessments using testlet compared to multiple choices scoring, with t value equal to 7.864 with significance <0.001. This difference can be seen from the average achieved when using scoring level. The average level can be seen in the figure 05.

To know the teachers' evaluation after test instrument of mathematic testlet model test workshop, the researcher was preparing the questionnaires consisting the statement. The gained information is based on the teachers' evaluation towards the instrument of Mathematics test instrument testlet model after joining the workshop for designing the teachers' test instrument related to their interest in applying the test instrument is 36.84% teachers said that "very good" and 63.16% teachers said "good" for applying the test instrument within the class assessment of elementary school. Based on the conversion table, the average for all items and responses related to the teachers' evaluation is 4.57. This result, if it is conversed goes to very good category, therefore it can be concluded that the average answer of teachers give a good value towards the development of Mathematics test instrument testlet model especially for the class assessment in elementary school. The common form of test used in the class assessment is elaboration test and dichtomic multiple-choice. Both test forms have strength and weaknesses. To overcome the weaknesses and combine between two forms of the test, the researcher developed Mathematics test instrument testlet model. The result of this research showed that the instrument of this model for a classroom assessment becomes the solution for teachers so they can check the test quickly as in a multiple-choice test, teachers also can know the students' weaknesses as it appears in the essay test so it can be utilized as the basic items to improve the instructional quality.

REFERENCES

Abrory, M. & Kartowagiran, B. (2014). Evaluasi implementasi kurikulum 2013 pada pembelajaran matematika SMP negeri kelas VII di kabupaten sleman. *Jurnal Evaluasi Pendidikan, 2 (1), 50-59*

Dikpora, (2015). *Data hasil ulangan semester bersama sekolah menengah pertama. Lombok Timur: Dikpora Kabupaten Lombok Timur tidak diterbitkan*

Edward H. Ip. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement 34(7) 467–482*

Griffin, P. & Nix, P. (1991). *Educational assessments and reporting.* Sidney: Harcout Brace Javanovich publisher.

Hambleton, R.K,; Swaminathan, H. & Rongers, H.J. (1991). *Fundamental of item respons theory.* Newbury Park, CA: Sage Publication Inc.

Haylock, D. and Thangata, F. (2007). *Key concept in teaching primary Mathematics*. London: SAGE Publication.

Humenberger, H. (1997). *Applicable mathematics education teaching mathematics*. Retrieved 25 July, 2014 from http://www.ex.ax.uk/cimt/ijmtl/hhempres.pdf)

Kartowagiran, B. & Jaedun, A. (2016). Model asessmen autenitk untuk menilai hasil belajar siswa sekolah menengah pertama (SMP): Implementasi asessmen autentik di SMP. *Jurnal Penelitian dan Evaluasi Pendidikan, 20 (2), 131-141*

Mardapi, D. (2004). *Penyusunan tes hasil belajar*. Yogyakarta: Universitas Negeri Yogyakarta

_____. (2008). *Teknik penyusunan instrumen tes dan nontes.* Yogyakarta Mitra Cendikia Press

Myford, C.M. & Wolfe, E.W. (2009). Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *Journal of Educational Measurement Winter, 46 (4), 371-389*

Paek, I.; Yon, H.; Wilson M. & Kang, T. (2009). Random parameter structure and the testlet model: extension of the rasch testlet model. *Journal of Applied Measurement. 10(4), 394-407*

Retnawati H.; Hadi S. & Nugraha, A.C. (2016). Vocational High School Teachers' Difficulties in Implementing the Assessment in Curriculum 2013 in Yogyakarta Province of Indonesia. *International Journal of Instruction, 9(1), 33-48*

Rohmawati (2013, Desember). *87 persen guru kesulitan dalam cara penilaian Kurikulum 2013,. www. unnes.ac.id › Berita edisi sabtu, 14 desember 2013.*

Russel, M.K. & Airasian, P.W. (2012). *Classroom assessment: concept and application 7th ed.* New York, NY.: McGraw-Hill

Susongko, P. (2010). Perbandingan keefektifan bentuk tes uraian dan testlet dengan penerapan graded response model (GRM). *Jurnal Penelitian dan Evaluasi Pendidikan Tahun 14, Nomor 2, 2010*

Tierney, R.D.,(2006). Changing practices: influences on classroom assessment. *Asessment in Education, 13(3), 239-264*

Van den Heuvel-Panhuizen (1996). *Assessment and realistic mathematics education*. Utrecht: CD-β press

Wainer, H & Kiely, G.L. (1987). Item cluster and computerized adaptive testing: a case for testlets. *Journal of Education Measurement, 24, 185-201*

Wainer, H,; Bradlow, E.T. & Wang, X. (2007). *Testlet response theory and its application.* New York, NY: Cambridge University Press

Yue, L. & Hong-Yun, H. (2012). When should we use testlet model? A comparison study of Bayesian testlet random-effects model and standard 2-PL Bayesian model. *Journal Acta Psychologica Sinica , 44 (2), 263−275.*