The characteristics of mathematical reasoning and proof test on indonesian high school students

IOP Conf. Series: Journal of Physics: Conf. Series **1200** (2019) 012007 doi:10.1088/1742-6596/1200/1/012007

The characteristics of mathematical reasoning and proof test on indonesian high school students

Y M Sari[1], B Kartowagiran[1], H Retnawati[1] and S Fiangga[2]

[1]Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Indonesia
[2]Department of Mathematics, Universitas Negeri Surabaya, Indonesia

E-mail: yurizka.melia@gmail.com

Abstract. Both reasoning and proof ability have been considered as two sides of the coin on which cannot be separated. The reasoning ability is developed by solving proof tasks on which the students' abilities in argumentation and justification are needed. However, the test items provided by the teachers tend to be routine problems by which the students' reasoning ability is limited. This study aims to test the instrument of mathematical reasoning and proof evaluation at the high school level. The instrument had been carried out on 61 high school students in East Java. QUEST was used for analyzing the students' answers. The result provides a value of INFIT MNSQ as 1,21 (around 1) and standard deviation 0,02 (around 0,0), which imply that the whole test fit with the PCM 1 PL model. The degree of difficulties of the developed high school reasoning and proof assessment instrument varies between -2.0 and +2,0, which imply that the instrument is categorized as good because the required criteria have been satisfied.

1. Introduction

The fact that the low level of students' mathematics achievement and mastery is undeniable. Most of the students tend to only remember the mathematics formula without any conceptual understanding on it which can be considered in knowing category based on revised Bloom's taxonomy. This claim is supported with the TIMSS (Trends in International Mathematics and Science Study) survey in 2015, especially in mathematics from which the survey shows that Indonesia is in the 45th rank out of 50 countries with 397 points [1]. In details, from TIMSS 2015, Indonesian students' reasoning ability is still low compared with knowing and application domain. Meanwhile, the PISA (Programme for International Students Assessment) 2015 result also suggests a parallel finding with TIMSS 2015 from which Indonesia is ranked on 63rd out of 69th participant countries [2]. These findings bring up a question on what actually happened in the Indonesian mathematics assessment.

Actually, in the recent Indonesian curriculum, a good mathematical ability required to be mastered by the students have been formulated. One of the required mathematical ability is the reasoning and proof ability for high school students. Based on revised Curriculum 2013, the recent official curriculum in Indonesia, a compulsory core ability that must be mastered by the students have been formulated which are arranging, reasoning, and presenting in concrete and abstract domains related to the development of knowledge independently from school and also be able to perform a scientific method [3]. Furthermore, in the high school assessment guide, there are a number of evaluation approaches to assess the mathematics achievement in which one of them is the assessment for learning [4]. This evaluation aims for improving mathematics teaching and learning in the class and mathematics ability

especially in reasoning skill [5]. This theoretical framework on mathematics assessment has been developed well but the facts relating to the implementation in the class are not as good as it is.

The implementation of assessment in mathematics teaching still uses routine problems. These routine problems tend to make students to remember and calculate using the remembered formula only [6]. Therefore, the students' mathematical paradigm or abstraction could not be developed well in the information transformation during the class. Halyoak & Morrison [7] stated that the students' thinking process evaluation through information transformation activity requires interaction between abstraction, reasoning, imagination, and problem-solving. The fact that there is a gap between mathematics assessment theory and practice in Indonesia is irrefutable especially in the issue of using routine problems in evaluating students' knowledge. However, the gap may be bridged by providing mathematics proof tasks in which the students' mathematics abstraction, reasoning, imagination, and problem-solving can be developed [8].

The reasoning and proof ability tasks cannot be considered separately. The students' ability in reasoning could be investigated from the students' argumentation used in deriving conclusion [9]. Moreover, proof in mathematics teaching and learning could improve the mathematical understanding as one of the instrument to reveal the mathematics understanding complexly, including students' mathematical reasoning [10]. However, mathematics assessment by using proof tasks to improve the reasoning ability in high school is still limited on mathematics induction problems [11]. Therefore, it can be concluded that the mathematical reasoning test tends to consider students' inductive reasoning only.

In mathematics education, mathematics assessment could be used in various purposes like providing information for helping teachers and students on understanding the material until the designing of national strategy in improving mathematics education. The aforementioned facts confirm that assessment is the key to promote effective education started from the improvement of classical assessment has been become the main focus of government education development effort [12]. Therefore, the stages in implementing evaluation must be prepared well and systematically starting from planning, implementing the procedure, and the evaluation instrument used.

In determining the students' ability in mathematics both individually and in a group, the teacher should consider various context from the mathematical ability performance, including the knowledge about mathematical aspects and mathematics disposition. Therefore, in collecting accurate information on the students, teachers may use the evaluation technique which is manifested as a test instrument. The developed test instrument could uncover the learning achievement that consists of academic proficiency and process standard, especially in mathematics education. NCTM [13] explained that there is process standard evaluated in mathematics which is problem-solving, reasoning, connection, and communication. Therefore, in this study, the developed test instrument focus on the revealed reasoning ability by using proof tasks in both empirical, deductive, and narrative.

In developing mathematics ability evaluation, especially in reasoning and proof ability, teachers should develop authentic assessment by which the students' input, process, and output components can be evaluated holistically [4]. The conducted evaluation would provide a portray of students' capacity, levels, and learning achievement in the ability of reasoning and proof ability that could produce instructional and accompanist impact from the mathematics instruction. Therefore, the result of the reasoning and proof ability evaluation developed by the teachers could provide important and crucial information in improving the instruction. In addition, the results could also be used as the reference on students' ability development through ages.

In obtaining a meaningful evaluation of the evaluation result analysis, teachers must consider the implementation of the evaluation especially on the quality of the developed test item. The analysis on the test item from which an item parameter is not categorized as good could not provide an accurate result from the conducted assessment. Besides that, the test problems should be really derived from the graduate competency standard, competency standard, basic competency, or indicators on the subject [14]. In this study, the development of the assessment instrument on reasoning and proof ability using Item Response Theory (IRT) where the test developer must construct the items based on the desired

design on the final format used. In the IRT approach, a big heterogenic calibrated sample is needed to estimate the required parameter [15]. Before conducting item analysis by using IRT approach which is partial credit model (PCM), initial analysis is required by using classical theory test to omit the items which have approximately nil score from the analysis result. This treatment is done in order to avoid data concentration. Besides, the researcher also conducted IRT calibration in estimating the item parameters and detecting compatibility data with the model used.

2. Methods

This research is a descriptive evaluative research with a quantitative approach, analyzing the result of the answer of high school student of class XI in Sidoarjo, East Java which has followed limited trials of reasoning and proof test. This study describes the characteristics of reasoning and proof instruments such as validity and reliability of tests. In addition, this study will also test the appropriate logistic model to determine the difficulty parameter and student ability estimation.

2.1. Population and sample (study group or participants)

The study was conducted in public high school in East Java, Indonesia in November – October 2017. The selection of schools was conducted by purposive sampling especially for schools that have implemented learning along with the assessment of reasoning and mathematical proof. Selection of public high school is expected to represent a large variation on the characteristics of learners such as family background, culture, ethnicity, religion, social, and economy.

The subjects of the trial include the high school students of XI-Science class who were present at the study site at the time of the test. Class X was not involved in the research because it has not obtained mathematical material about reasoning and verification in accordance with the current 2013 revised curriculum. While the students from class XII-Science also was not involved as a research subject because it is facing the National Exam in 2018. The testing of the assessment tool had been done involving 61 respondents.

2.2. Data collection method(s)/technique(s)/tool(s)

Research data in the form of scores on the results of reasoning tests and mathematical proof. Scores of test results in the form of data polytomous four categories. The data collection instrument uses a mathematical reasoning and mathematical probe which consists of 5 items. It is based on categories of reasoning and mathematical proofs of making or investigating conjectures, developing or evaluating arguments, and correcting false arguments. Each category contains one good question related to the number or geometry material that the learner has learned. Each test device is constructed taking into consideration the representation of every aspect of reasoning and mathematical proof that will be measured.

2.3. Data Analysis

Analysis of this research data used Partial Credit Model 1 PL (Parameter Logistic) for the testing fit item of high school reasoning and mathematical proof test. Basic considerations were used, the first that PCM as an extension Rasch Model which is a 1-PL model used a sample that is not so large compared to if doing a calibration of data polytomous using model 2-PL or 3-PL [16]. Secondly, that the response characteristics of the reasoning and mathematical proof items follow the PCM.

Data analysis was performed on several aspects, namely: (1) instrument item match, (2) reliability, (3) item characteristic curve (ICC), (4) difficulty index, and (5) information function and SEM. Testing of goodness of fit for overall test and test case (case / person) as a whole was developed by Adam and Khoo [17] based on the mean value of INFIT Mean of Square (Mean INFITMNSQ) and its standard deviation or observed the mean value of INFIT t (Mean INFIT t) along with its standard deviation. If the INFIT MNSQ average is about 1.0 and the standard deviation is 0.0 or the INFIT t rate is close to 0.0 and the standard deviation is 1.0, then the whole test is fit with the model.

The match of the item with the model is known to fit the item and the test participant follows the rule that the characteristic curve (ICC) item will be flat if the INFIT MNSQ size for the item or e is greater

than the logit unit> 1.30 or <0.77. This state of the distribution graph forms a platicurtic curve and no longer forms a leptokurtic curve [18]. Therefore, an item or person is declared fit with a model with a limit of INFIT MNSQ ranges from 0.77 to 1.30 [17]. In this case, the t value range is ± 2 (rounding ± 1.96) if the error or alpha level is 5% [18]. Item is said to be good if the index of difficulty is more than -2.0 or less than 2.0 [19]. Based on the information function and SEM, it can be seen that this test is suitable for students with low, medium or high ability.

3. Result and Discussion

Student response data in this research was polytomous scored with 5 categories which are category 4, 3, 2, 1, and 0. The criteria for achieving each category level if fulfilling the requirements in the Table 1.

Table 1. The Criteria for Student Achievement

Level	Description
4	Providing answers (formal and narrative) in the form of valid arguments, the steps used are complete, the explanation is acceptable (justification is correct), using symbols and math language well
3	Giving answers (formal and narrative) is not perfect, the steps used are mostly incomplete, the explanation is less acceptable, requires some justification
2	Give examples (either math pictures or formulas) and there is justification
1	Give an example (either a picture or a mathematical formula) but there is no justification
0	No response

3.1. Goodness of Fit

Testing of the goodness of fit polytomous data 5 categories analyzed according to Partial Credit Model (PCM) is done for the overall test or each item. The overall test fit test uses the rules developed by Adam and Khoo [17] based on the mean value of INFIT Mean of Square (Mean INFITMNSQ) and its standard deviation or observed the mean value of INFIT t (Mean INFIT t) and its standard deviation. If the INFIT MNSQ average is about 1.0 and the standard deviation is 0.0 or the INFIT t rate is close to 0.0 and the standard deviation is 1.0, then the whole test is fit with the PCM 1 PL model.

The results of the test match analysis seen from the INFIT parameter for Mean Square (MNSQ) indicated that the instrument of assessment of mathematical reasoning and proof of mathematics meets the fit statistic criteria according to PCM which is presented in Table 2. The result of the 5 test analysis has INFIT MNSQ 1, 21 (about 1) and the standard deviation of 0.02 (about 0.0), then the entire fit test with PCM 1 PL model.

Table 2. Item Estimation Results and TPP Participants according to PCM 1-PL

Description	Item estimation	Testee estimation
Mean dan Standar deviation	-0,29 ± 0,51	0,19 ± 0,01
Standardize Mean dan standar deviation	0,00 ± 0,50	0,11 ± 0,00
Reliability	0,89	
Mean and standard deviation INFIT MNSQ	1,21 ± 0,02	1,21 ± 0,07
Mean and standard deviation OUTFIT MNSQ	1,21 ± 0,02	1,21 ± 0,12

Mean and standard deviation INFIT t	0,84 ± 1,47	0,25 ± 1,08
Mean and standard deviation OUTFIT t	0,27 ± 0,48	0,08 ± 0,61

The fit-setting test of each item in the model follows the rules of Adam and Khoo [17] ie a fit item on the model if INFIT MNSQ values are between 0.77 and 1.30. With item acceptance limit using INFIT MNSQ or fit according to the model (between 0.77 and 1.30) and using INFIT t with the range of -2.0 to 2.0, the items matching fit the goodness of fit. MNSQ INFIT Value Reasoning and Proof Tests (TPP) between 0.98 to 1.05. With item acceptance limit using INFIT MNSQ or fit according to model (between 0.77 up to 1.30), then all items are 5 items fit with PCM. Based on the analysis results obtained TPP test reliability is 0.89. This reliability value belongs to the high category.

4. Item Characteristic Curve and Index Difficulty (b)

The characteristic of the item is indicated by the item characteristic curve (ICC) and the difficulty index. In Figure 2 there is an example of ICC for item 1, which can be explained that: (a) score 0 (category 1) is mostly obtained by students with very low abilities ($\theta = -3$), (b) score 1 (category 2) obtained by students with low ability ($\theta = -2$), (c) score 2 (category 3) mostly obtained students with moderate abilities ($\theta = 0.1$), (d) score 3 (category 4) ($\theta = 1.6$), and d) score 4 (category 5) mostly obtained by students with very high ability ($\theta = 2.7$).

An index of difficulty or difficulty level (b) for a score of 0 (b0), a score of 1 (b1), a score of 2 (b2), a score of 3 (b3), a score of 4 (b4), and a roughness as difficulty. Based on the analysis, the difficulty of items lies between -0.08 to 1.369. Item is said to be good if the index of difficulty is more than -2.0 or less than 2.0 (So, based on difficulty, all items of 5 items are all good.

Table 3. Item Difficulty Level

Aspect	Sub	Difficulty Level	Category			
			1	2	3	4
Conjecture	Making	-0.052	-2.75	-0.90	-0.10	2.37
	Investigating	0.447	-1.38	-0.39	0.20	4.19
Argument	Developing	-0.082	-2.69	-1.31	-0.40	2.62
	Evaluating	1.369	0.03	1.37	4.12	
Correct a Mistake	Correct a mistake	-0.713	-4.22	-1.86	-0.95	0.34

Table 3 explains the extent of difficulty in the sub-aspect and instrument aspects of each category in PCM. Based on Table 3 also can be known the difficulty level on each sub-aspect and instrument aspect for each category in PCM.

5. Results Measurement Response Learners

Testing Measurement of reasoning and mathematical proof of high school mathematics subjects followed by 61 respondents who are in class XI Science. The distribution of scores of all learners ranges from 0.1 to 0.19 in logit scale between -4 to +4. The students' reasoning and proof-of-math skills can be estimated using the average percentage of students responding correctly to every measured aspect and sub-aspect. Based on Table 4, it is known that students' responses are more dominant in categories 1 and 2 and only a small percentage of respondents are able to achieve category 3. In other words, the ability of reasoning and proofing of students still exist that have not been satisfactory.

The instrument of reasoning and proof ability performance evaluation on high school mathematics subjects consists of 5 good items which are designing conjecture, conjecture investigation, argumentation development, argument evaluation, and mistake correction, fulfil the valid criteria. The

validity criteria used in this study is empiric validity proofed by the goodness of fit based on partial credit model (PCM). The value of INFIT MNSQ 1,21 (around 1) and standard deviation 0,02 (around 0,0), so the whole test fit with the PCM 1 PL model [17]. This result also suggests that the test items contain high order thinking like reasoning and proof that invites the students to think deeply about the subject material [20].

Supporting factors that maintain the validity criteria of the test instrument can be categorized as (1) the developed test items is derived from the indicators obtained from the measured aspects and (2) high school mathematics reasoning and proof instrument had been validated through professional judgment involving several educational practitioners. Third, the respondents solve the tasks in the instrument by using their real skill in the teachers' observation in their school.

The total information function of the test is relatively high for the score from -2,0 to +2,0. This result indicates that the developed instrument has a relatively high reliability because it is comprised from a high information function items developed based on the tested students' ability [19]. The index difficulty of the reasoning and proof evaluation instrument in high school mathematics which was developed in this study is varied from -2.0 to +2,0, is considered in a good category because it has already fulfilled the required criteria. The test item with difficulty index -2,0 is categorized as easy level. Meanwhile, the test item with index difficulty +2,0 is categorized as hard level. Therefore, the developed instrument fulfill the requirement good category [19].

6. Conclusion

The result of the reasoning and proof ability measurement on SMA Negeri (State high school) in Sidoarjo, East Java, followed by 61 respondent provide score distribution between 0.10 and 0,19 in logit scale between -2 and +42. The average of the measurement result is 0,13±0,015. Based on the average score on the students reasoning and proof test, the students perform an average ability in mathematics reasoning and proof. The estimated students' reasoning and proof ability in every aspects and sub-aspects measured provide the information that the students' responses are dominantly categorized in 1 and 2 with only a few numbers of respondents had achieved category 3. This situation indicates that the high school students mostly are categorized in the medium level of reasoning and proof ability. Based on the analysis result, it is suggested that a further study is required in using polytomous data analysis based on generalized partial credit model (GPCM 3PL).

References

[1] Mullis I V, Martin M O, Foy P, & Hooper M 2016 *TIMSS 2015: International Results in Mathematics* (TIMSS & PIRLS: Boston College)
[2] OECD 2016 *PISA 2015 Results in Focus* (New York: Columbia University)
[3] Kemendikbud 2013 *Permendikbud No. 65 Tentang Standar Proses Pendidikan Dasar dan Menengah* (BSNP: Jakarta)
[4] Dikdasmen. (2015) *Panduan penilaian untuk sekolah menengah atas*. Jakarta: Kemendikbud
[5] Komatsu K, Jones K, Ikeda T, & Narazaki A 2017 Proof validation and modification in secondary school geometry *J. of Math. Behavior* **47** 1–15
[6] Fiangga S 2014 *Proc. Int. Conf. on Research, Implementation, and Education of Mathematics and Sciences* (Yogyakarta: Universitas Negeri Yogyakarta) pp 453–460
[7] Halyoak K J and Morisson R G 2005 *Thinking and reasoning: a reader's guide. Dalam The Cambridge Handbook of Thinking and Reasoning* (New York: Cambridge University Press) pp 1-9
[8] Blanton M L and Stylianou D A 2014 Understanding the role of transactive reasoning in classroom discourse as students learn to construct proofs. *J. of Math. Behavior* **34** 76–98
[9] Charlesworth R 2005 Prekindergarten mathematics: Connecting with national standards *Early Childhood Ed. J.* **32(4)** 229–36
[10] Hanna G 2000 Proof, explanation, and exploration: An overview *Ed. Stud. in Math*. **44** 5–23

[11] Imamoglu Y and Togrol A Y 2015 Proof construction and evaluation practices of prospective mathematics educator *Euro. J. of Sci. and Math. Ed.* **3 (2)** 130-44
[12] Christoforidou M, Kyriakides L, Antoniou P and Creemers B P M 2014 Studies in educational evaluation searching for stages of teacher's skills in assessment. *Studies in Educational Evaluation* **40** 1–11
[13] NCTM 2000 *Principles and standards for school mathematics* (Reston VA: NCTM)
[14] Westen D and Rosenthal R 2003 Quantifying construct validity. *J. of Personality and Social Psychology* **84 (3)** 608-18
[15] Oriondo L L and Dallo A 1998 *Evaluating educational outcomes (test, measurement, and evaluation)* 5th ed. (Quezon City: REX Printing Company)
[16] Keeves J P and Masters G N 1999 *Advances in Measurement in Educational Research and Assessment* (Amsterdam: Pergamon, An imprint of Elsevier Science) pp 1-22
[17] Adams R J and Khoo S T 1996 *Quest: The interactive test analysis system version 2.1*. (Victoria: The Australian Council for Educational Research)
[18] Keeves J P and Alagumalai 1999 New Approach to measurement *Advances in Measurement in Educational Research and Assessment* ed G N Masters & J P Keeves (Amsterdam: Pergamon, An imprint of Elsevier Science) pp 23-42
[19] Hambleton R K Swaminathan H and Rongers H J 1991 *Fundamental of item response Theory* (Newbury Park, CA: Sage Publication Inc.)
[20] Ball D and Bass H 2003 Making mathematics reasonable in school *A research companion to principles and standards for school mathematics* ed J Killpatrick, G Martin and D Schifter (Reston, VA: NCTM) pp 27-44