Development of Higher Order Thinking Skills (HOTs) Assessment Instruments in Mathematics Subjects for Class VIII Middle Schools

S Bakhri1 and R Rosnawati2

1 Mathematics Education of Postgraduate Program, Yogyakarta State University, Colombo Street, Yogyakarta, Indonesia

2 Mathematics Education Departement, Faculty of Mathematics and Science, Yogyakarta State University, Colombo Street, Yogyakarta, Indonesia

syaifulloh.bakhri2016@student.uny.ac.id

**Abstract.** The objectives of this study are (1) to describe good quality *HOTs* assessment instruments in the mathematics subjects to SMP VIII, and (2) to describe *HOTs'* ability of SMP students in class VIII. This research was a development study adapted from the Mardapi development model, which includes the following steps: (1) compiling test specifications, (2) writing test items, (3) reviewing test items, (4) test trials, (5) analyzing test items, (6) revising test items, (7) compiling test items, (8) testing items, (9) interpreting test results. The research was located in SMP N 1 Bantul and SMP N 2 Bantul. The research instrument was in the form of Mathematics *HOTs* questions for VIII grade junior high school. The quality of instrument products was reviewed from classical theories and Rasch models. Limited trials were conducted on 123 VIII grade students of junior high school. Field trials were conducted on 420 VIII grade students of junior high school. The research was conducted in the even semester of the 2018/2019 school year from May to June 2019. In the classical theory, the instrument validity and reliability are described, while the Rasch model was described as the fit and level of item difficulty. Analysis techniques using the Quest application. In the analysis phase, the items were analyzed with the QUEST application to find out (1) the suitability of instrument items (fit MNSQ), (2) the level of difficulty of the items. The results showed that (1) the *HOTs* assessment instruments produced were *HOTs* test questions consisting of 29 multiple choice questions and 13 description items that were suitable for use with the Aikens V index on the validity of construction aspects, language aspects, and content aspects of 0,45 to 0.90, the content reliability index meets the specified reliability criteria. Based on the classical theory, the validity index of the limited trial extracts five items are invalid and all items are valid field trials, the reliability indexes of the limited trial instruments and field trials meet the criteria predetermined. Based on the theory of response items limited trials five items do not fit the model and three items that do not meet the level of item difficulty, field trials of all items, fit the PCM model as evidenced by the *MNSQ infit* value from 0.77 to 1.22, and meet the difficulty item level (b) from -1.95 to 2.82, so the instrument meets the valid and reliable criteria. Thus it can be concluded that item analysis according to classical test theory will be accurate when using small sample sizes and according to theory the response of items will be accurate when using large sample sizes. (2) the ability of students of class VIII *HOT* (θ) is in the high category with a level of ability (θ) from -1.04 to 2.82.

1. Introduction

The success of education, whose main objective is to improve human resources, is influenced by various factors. One of the factors that influence this success is the ability of teachers to carry out and take advantage of assessments, evaluation of learning processes and outcomes. This ability is needed to determine whether the learning objectives set in the curriculum have been achieved [4]. Achievement of learning objectives can be seen from student achievement as measured by assessment. Therefore, the position of assessment of learning outcomes is important in achieving learning objectives.

According to Nitko and Brookhart [11], “assessment is a broad term defined as a process for obtaining information that is used for making decisions about students; curricula, programs, and schools; and educational policy”. In line with this, Van de Walle stated that assessment principles and standards emphasize two main ideas, namely that assessment should improve student learning and assessment is a valuable tool for making teaching decisions [4]. Furthermore, according to McMillan (2000) [3] suggests that "assessment is inherently a process of professional judgment." Furthermore, Cecil (2009: 3) [11] expressly states that "assessment is any systematic procedure for collecting information that can be used to make inferences ....".

This is in line with the opinion [1] which states: Assessment is an important aspect and should be the main concern in determining the direction of a program or national education system. Thus, HOTs elements in the subject of mathematics assessment need to be emphasized since it can measure teachers understanding on the implementation of HOTs in mathematics classroom. This is supported by the opinion of [5]which states that: The quality of learning outcomes assessment instruments will influence directly in the achievement of student learning outcomes. Therefore, the position of learning outcomes assessment instrument is strategic for teachers and schools in decision making related to learning outcomes achievement including high order thinking skills.

Related to high-order thinking skills, Brookhart argues that higher-order thinking skills (HOTs) include logic and reasoning skills, analysis, evaluation, and creation, problem solving. solving), and decision making (judgment) [9]. This is also supported by [10] which states that "the next five standards addres the processes of problem solving, reasoning and proof, connections, communication, and representation".

This proves that HOT is closely related to problem-solving abilities, but the problem-solving abilities of Indonesian students are low [12]**.** These findings are in line with the results of the study of The Third International Mathematic and Science Study Repeat (TIMSS-R) which revealed that the mathematical ability of Indonesian junior high school students for non-routine questions was very weak, but relatively good at solving fact and procedure questions. Based on the results of the TIMSS-R study and the results of research conducted by Herawati, it appears that for math problems that require higher-order thinking skills, Indonesian students are still far below the international average. The ability to solve problems, critical thinking skills, and creative thinking skills of junior high school students in Indonesia is still low, so that students are weak in solving non-routine questions [8]. This is in line with the findings of [2] which is shown by the Cronbach's Alpha value in the question package C which is not reliable, namely 0.488. The results of the instrument trial concluded that the students' higher order thinking (HOT) ability in mathematics was not good. This can be seen from the average value of the trial results of 26.38 on a scale of 100. This is in line with the opinion of [4] who stated that high-order thinking skills among students are still low.

In a study conducted by [4], the analysis of test quality related to HOTs was carried out using classical test theory which resulted in students' abilities which affected the results of the analysis. Furthermore, in research conducted by [7], that students are not continuously trained to have higher-order thinking skills, therefore, it is necessary if teachers optimize assessment techniques in the form of tests that can be used to hone students' thinking skills. , in this case HOTs. Other findings data show that teachers understand that there is a revision in K 2013, including developing HOTs, including developing assessment instruments, but they have difficulty in formulating indicators in HOTs into assessment instruments, so it is necessary to develop HOTs assessment instruments that are well qualified in the eyes. mathematics class VIII SMP.

1. Research Methods

This research is a development research that aims to develop a test instrument product in the form of an instrument to measure higher order thinking skills (HOTs) in Mathematics at Class VIII Junior High School. The test instrument developed in this study used the test development model developed by Mardapi, namely: (1) compiling test specifications, (2) writing tests, (3) examining tests, (4) conducting trials, (5) analyzing test items , (6) improving the test, (7) assembling the test, (8) conducting the test, and (9) interpreting the test results. The nine stages of developing the test instrument are carried out systematically so that a good test instrument product is produced. The development procedure is a development step based on the Mardapi development model used below [6].

|  |
| --- |
|  |
| **Figure 1.** Mardapi Test Instrument Preparation Procedur | |

The trial design was carried out by giving a series of tests in the form of HOTs items that had been compiled to a number of samples of students. The test trial activity is carried out in the classroom, participants are asked to work independently and with a predetermined duration of time. During the trial activity, a teacher is the supervisor. The collected answers are then analyzed using the help of the Quest program. The analysis results in the form of item characteristics, items with good characteristics are used and items with poor characteristics are aborted. Items are aborted if they have very bad characteristics. The trial subjects consisted of VIII grade students in junior high schools in Yogyakarta Special Region. The population in the study consisted of all junior high school students in DIY. This type of sampling using proportionate random sampling. This technique determines the number of samples based on the representation of the sample to the population to be studied. According to Wagiran, the random sampling technique provides equal opportunities for each member of the population to be used as research samples [6].Data collection was carried out during the process of preparing the assessment instrument as well as in the process of assessing the results of student responses, including through:Test the appropriateness of the assessment instrument developed based on expert judgment that proves the validity and reliability of the instrument, tried out the assessment instruments developed in class VIII junior high school students and saw the scores of the responses of students' answers to determine the feasibility of the instruments developed. The data analysis technique in this study was to obtain an assessment instrument that met the proper criteria so that a quality product was obtained, among others through: Classical Theory and IRT.

1. The Research Result

The first stage is to compile a test specification that has several steps that need to be taken, including determining the goal of developing the test, creating a test grid, choosing the test form, and finally determining the length of the test. The making of the HOTs test grid is based on the Core and Basic Competencies in the 2013 Curriculum in mathematics at the VIII grade of junior high school level. Five experts were assessed by five experts. Validation is carried out to see the initial product before the limited trial which aims to obtain validity assessments, input and suggestions for improvement of HOTs assessment instruments. The results of the assessment by five experts were then analyzed with the Aiken's V formula. Based on the results obtained from the analysis using the V Aiken index, and looking at the validation level table by Suharsimi Arikunto, it can be concluded that the instrument developed had a V Aiken index value of 0.45. up to 0.90.

To determine the reliability value of the contents of the instrument analyzed by the formula C. Hoyt. The steps are: 1) Calculating the Total Sum of Squares (JKT), 2) Calculating the Sum of Squares of Items (JKi), 3) Calculating the Sum of the Squares of the Subjects (JKs), 4) Calculating the Sum of the Squares of the Interactions (JK int). Based on the results of the assessment of the five validators that have been analyzed, the average results of the content reliability of each factor are as follows. First, the construction factor = 0.928. Second, the language factor = 0.934. Third, the content factor = 0.937. The mean result of the content reliability index for all factors was 0.933. The average result of the content reliability index for each factor was 0.933, indicating that the content reliability index for each factor was higher or above the specified reliability criteria, namely 0.6. Thus it can be concluded that based on the validator's assessment, the instrument items developed can be categorized as meeting content reliability. With the fulfillment of the content reliability value, it shows that the validator's assessment of the instrument items was not a significant difference between the five raters. This also indicates that there is an element of significant similarity between assessors or validators to the items of the assessment instrument. By producing a good content validity index and content reliability, the instrument assessment process is complete. The product results of the assessment instrument for limited trials.

The limited trial was carried out by involving 123 junior high school students in grade VIII. The limited trial was carried out for two days, each to solve the questions on packages A and B. At the instrument testing stage, quantitative data were obtained. Furthermore, one of the unidimensional assumptions can be proven by using factor analysis, to see the eigenvalues ​​of the inter-item covariance variance matrix. Data analysis with factor analysis was preceded by an analysis of the adequacy of the sample. This study proved unidimensional assumptions on the test taker's response data to HOTs questions in mathematics at grade VIII junior high school. Based on the analysis of the adequacy of the sample, it shows KMO> 0.5, the Chi-squared values ​​in the Bartlet test are 2027.409 and 2621,200 with 561 and 120 degrees of freedom and p values ​​less than 0.05. These results indicate that the sample size of 123 used in this study is sufficient.

Based on the graph above, it can be seen that in general the questions used have only one factor component, so it can be concluded that there is only one dominant factor for HOTs questions. So for the unidimensional conditions are met, so that the other two assumptions can be represented. Furthermore, the validity of the construct and the reliability of the experimental test results were sought with classical theory analysis using the Quest program. Based on the results of testing the validity of the construct, items are declared valid if they have an infit value of MNSQ in the range of 0.77 to 1.30 (Bashooir, & Supahar, 2018: 225). The analysis performed using the help of the Quest program shows that all 45 items have an MNSQ infit that is in the range 0.77 to 1.30 which is declared valid, while 5 items are declared invalid.

|  |  |
| --- | --- |
|  | |
| **Figure 2.** The results of the MNSQ infit distribution after field trials were carried out with 420 students who were analyzed using the Quest program. According to Bambang Subali in 2011, the item is said to be fit in the model if the MNSQ infit value is from 0.77 to 1.30. If more or less, the item will be rejected. Based on these results, all item items were declared fit with the PCM Model. | | |
|  | | **Figure 3.** The distribution of difficulty levels (b) is seen from the results of the Quest program. From the results of the analysis of the level of difficulty, it can be concluded that all items fall into the good category because they have a difficulty level from -2.00 to +2.00 (-2.00 ≤ b ≤ +2.00). The easiest item of questions is question number 17 with a difficulty level of -1.95 and the most difficult item is number 22 and 34 with a difficulty level of +1.98. | | |

Based on the results of item analysis according to classical theory and the rasch model carried out with the help of the Quest program in the field trial, it shows that the item analysis according to the classical test theory in field trials gives results in the form of the same number of good items according to the Rasch model. The results of the analysis in field trials turned out to be contrary to the results of the analysis in limited trials. Thus it can be concluded that the item analysis according to classical test theory will be accurate when using a small sample size and according to the item response theory will be accurate when using a large sample size. Based on the results of the data analysis carried out, it can be seen that the question instrument has validity, reliability, fit with the PCM model, and a level of difficulty that meets the appropriate criteria to be used to measure the HOT ability of class VIII junior high school students.

Measurements were carried out using the VIII grade mathematics HOTs test instrument in the form of multiple choice and a description which aims to describe the ability of students in higher-order thinking in mathematics. Measurement activities were carried out in two class VIII junior high schools in the Yogyakarta Special Region. After the question instruments were done by students, then the data on the students' work were analyzed with the Quest program. The output of the Quest program is an estimate of the HOT ability of grade VIII junior high school students (θ). The distribution of HOT abilities of class VIII junior high school students can be seen in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.** Distribution of HOT ability scores for VIII grade junior high school students | | | |
| Value θ | Level | N students | Persentage |
| +1,00 < θ | Very high | 138 | 32,86% |
| +0,30 < θ ≤ +1,00 | High | 169 | 40,24% |
| -0,30 < θ ≤ +0,30 | Moderate | 67 | 15,95% |
| -1,00 < θ ≤ -0,30 | Low | 45 | 10,71% |
| θ < -1,00 | Very low | 1 | 0,24% |

The θ value ranges from -1.04 to 2.82. With the results of this analysis according to the conversion table of quantitative to qualitative values ​​from Syaifudin Azwar, the VIII grade junior high school students fall into the high category in HOT ability. Here are some examples of questions from each aspect.

1. Conclution

The assessment instrument developed was a multiple choice test with four alternative answers as many as 29 items and a description of 13 items. The measured cognitive domain includes analyzing (C4), evaluating (C5), and creating (C6). The assessment instrument developed was a multiple choice test and a description to measure the HOT ability of class VIII junior high school students with good qualifications with the following criteria: The assessment instrument developed in the form of a multiple choice test and the description has met the validity of the content obtained from the assessment of five experts with the V Aikens index in the range 0.45 to 0.90, and has met the validity of the instrument construct. The assessment instrument developed in the form of a multiple choice test and the description has met the requirements for content reliability and instrument reliability. The assessment instrument developed in the form of a multiple choice test and the description has met the fit requirements with the PCM model as evidenced by the MNSQ infit value from 0.77 to 1.22. The assessment instrument developed was a multiple choice test and the description had met the difficulty level requirements of item (b) from -1.95 to 1.98. The HOT ability of grade VIII junior high school students is in the high category with θ values ​​ranging from -1.04 to 2.82. The classical theory analysis is based on validity and reliability, while the rasch model analysis is based on threshold and fit  
 Item analysis in limited trials according to classical theory gives results in the form of the number of good items more than according to the Rasch model, while item analysis in field trials according to classical theory gives results in the form of as many good items as according to the Rasch model, thus It can be concluded that the item analysis according to the classical test theory will be accurate when using a small sample size and according to the item response theory will be accurate when using a large sample size.

References

[1]Abdullah A H, Mokhtar M, Halim, N D A, Ali D F, Tahir L M & Kohar U H A 2017 Mathematics teachers’ level of knowledge and practice on the implementation of higher-order thinking skills (HOTs) *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(1). https://doi.org/10.12973/eurasia.2017.00601a

[2]Arifin Z & Retnawati H 2017 Pengembangan instrumen pengukur higher order thinking skills matematika siswa SMA kelas X Tesis Magister tidak diterbitkan Universitas Negeri Yogyakarta

[3]Brookhart S M 2011 Educational Assessment Knowledge and Skills for Teachers, *30*(1), 3–12

[4]Budiman A & Jailani 2014 Pengembangan instrumen asesmen higher order thinking skill (HOTs) pada mata pelajaran Matematika SMP kelas VIII semester 1 *Jurnal Riset Pendidikan Matematika*, *1*(2), 139–151 Retrieved from https://journal.uny.ac.id/index.php/jrpm/article/view/2671/2224

[5]Budiman A & Jailani 2015 Developing an Assessment Instrument of Higher Order Thinking Skill (HOTs) in Mathematics for Junior High School Grade VIII Semester 1 *Proceeding of International Conference On Research, Implementation And Education Of Mathematics And Sciences 2015*, (May), 17–19

[6]Firdausa A R. & Haryanto 2019 Pengembangan instrumen tes high order thinking skills (HOTs) boga dasar untuk siswa SMK program keahlian tata boga Tesis Magister tidak diterbitkan Universitas Negeri Yogyakarta

[7]Hanifah N 2019 Pengembangan Instrumen Penilaian Higher Order Thinking Skill (HOTs) di Sekolah Dasar *Current Research in Education: Conference Series Journal*, *1(1)*, 1–8

[8]Iskandar J & Riyanti R 2015 Upaya Meningkatkan Kemampuan Berpikir Kreatif Siswa SMP dengan Pendekatan Matematika Realistik Indonesia *Seminar Nasional Matematika dan Pendidikan Matematika UNY*, 861–866

[9]Kurniati D, Harimukti R. & Jamil N A 2016 Kemampuan berpikir tingkat tinggi siswa SMP di kabupaten Jember dalam menyelesaikan soal berstandar PISA *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*, 142–155

[10]National Council Of Teachers Of Mathematics 2000 Principles and Standards for School Mathematics *School Science and Mathematics*, *47*(8), 868–279. https://doi.org/10.1111/j.1949-8594.2001.tb17957.x

[11]Sudarwan R E & Retnawati H 2015 Pengembangan perangkat assessment pembelajaran matematika pokok bahasan geometri dan pengukuran SMP/MTs *Jurnal Riset Pendidikan Matematika*, *2*, 251–261

[12]Widodo T & Kadarwati S 2013 Higher order thinking berbasis pemecahan masalah untuk meningkatkan hasil belajar berorientasi pembentukan karakter siswa *Jurnal Cakrawala Pendidikan*, *5*(1), 161–171. https://doi.org/10.21831/CP.V5I1.1269