**M – 12**

# DEVELOPING AN ASSESSMENT INSTRUMENT OF HIGHER ORDER THINKING SKILL (HOTS) IN MATHEMATICS FOR JUNIOR HIGH SCHOOL GRADE VIII SEMESTER 1

**Agus Budiman, Jailani**
SMP Negeri 2 Mandiraja, Yogyakarta State University
agusbudimath@yahoo.com, jailani@uny.ac.id

**Abstract**

This research aims to produce a valid and reliable mathematics assessment instrument in the form of HOTS test items, and describe the quality of HOTS test items to measure the high order thinking skill of grade VIII SMP students. This study was a research and development study adapting Borg & Gall's development model, including the following steps: research and information collection, planning, the early product development, limited try out, revising the early product, field try out, and revising the final product. The research's result shows that the HOTS assessment instrument in the form of HOTS test items consists of 24 multiple – choice test items and 19 essay test items, based on the judgement of the materials, construction, and language is valid and appropriate to be used. The reliability coefficients of the instrument are 0.713 for the multiple choice items, and 0.920 for essays. The multiple choice items has the average of item difficulty 0.406 (average), the average of item discrimination 0.330 (good), and the distractors function well. The essay test items has the average of item difficulty 0.373 (average) and the average of item discrimination 0.508 (good).

**Keywords:** development, assessment instrument, Higher Order Thinking Skills (HOTS), mathematics in the junior high school

## Introduction

Principle and assessment standards emphasize two main ideas that is assessment should improve learning of students and assessment is a valuable tool for teaching decision making (Van de Walle, 2007, p.78). The assessment is not only a data collection of students, but also related to data processing in order to obtain an overview of the process and the students learning outcomes. The assessment is not merely giving test then finished, but teacher also have to follow up the learning. In carrying out the assessment, teacher needs assessment instruments in good test items for testing the abilities of cognitive, affective, and psychomotor.

Assessment is an important activity in mathematics. Assessment can provide constructive feedback for teachers as well as students. Assessment results can also provide motivation for students to achieve better. Even assessment can influence the learning behavior because students tend to direct their learning activities towards the assessment that conducted by teacher. The quality of learning outcomes assessment instruments will influence directly in 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.

Students' lackluster in higher order thinking skills has attracted educators and Mathematics education researchers as implied in the statement Henningsen & Stein (1997, p.524) "many discussion and concern have been focused on limitations in students' conceptual understanding as well as on their thinking, reasoning, and problem solving skills in

mathematics". In Indonesia, students' lack of knowledge in mathematics, are always a topic of conversation in the community. Students were obstacled to use their knowledge of mathematics in everyday life, not even able to use skills solve if given a slightly different question from what is learned. Results of a survey on student achievement which were undertaken internationally, showed that Indonesian students were still far below the average. It was presented by research Trends in International Mathematics and Science Study (TIMSS) once every four years to measures the ability of students of class VIII junior high.

Mullis, et al. (2012, p.56) state that the achievements of TIMSS in 2007 and 2011 showed learning achievement scores of students class VIII junior high (eight grade), 397 and 386 consecutive (scale 0 to 800) with an average score of 500. The ability of junior high students class VIII Indonesia were below average. The results did not show much change on each of its participation. The result of low achievement TIMSS certainly is caused by several factors. One of the factors, was because students in Indonesia were less trained in solving problems of contextual, demanding reasoning, argumentation and creativity in completing it, where such questions are characteristic problems of TIMSS. This is in accordance with *Kemdikbud (*2013, p.2*)* which stated that the low student achievement Indonesian was caused by large number of test material in TIMSS are not found in the curriculum of Indonesia.

Mullis, et al. (2012, p.30) states on TIMSS 2011 assessment domain in junior high students class VIII includes content domains and cognitive domains, each of which consists of several domains. Domain content in line with the object (content) to the content standards in Mathematics junior, namely: number, algebra, geometry, data and chance. Cognitive domain consists of knowledge (knowing), application (applying), and reasoning. Problems of mathematics developed by TIMSS demanding students to think low level to a high level. Problems with the demands higher level thinking associated with cognitive reasoning which among others include the ability to find a conjecture, analysis, generalization, connections, synthesis, not routine problem solving, and the justification or proof.

Characteristics of HOTS expressed Resnick (1987, p.3) which are non-algorithmic, complex, multiple solutions (many solutions), involves a variety of decision making and interpretation, application of multiple criteria (many criteria), and effortful (requires a lot of effort). Conklin (2012, p.14) states the following characteristics of HOTS: *"characteristics of higher-order thinking skills: higher order thinking skills encompass both critical thinking and creative thinking"*. Critical and creative thinking are two very basic human capabilities because both can encourage someone to always look at every problem faced critically and trying to find the answer creatively in order to obtain a new thing better and beneficial for life.

Questions or tasks that trigger the students to think analytical, evaluative, and creative can practice students in higher order thinking skills. Associated with these cognitive aspects, (NCTM, 2000, p.7) suggests *"the next five Standards addres the proceses of problem solving, reasoning and proof, connections, communication, and representation"*. These skills include the high level of mathematical thinking. The fact that happened in school, the questions tend to be more testing of memory less practice HOTS or higher order thinking skills of students, in case some of Competency Standards (SK) and the Basic Competency (KD) in mathematics can be developed HOTS questions.

Higher order thinking skills enhancement has become one of priorities mathematics lessons in the school. Students in junior high school/MTs should begin to be practiced to higher order thinking according to the age, it is accordance with BSNP (2006, p.139) stated that Mathematics is given to all students to equip them with the ability to think logically, analytically, systematically, critical, and creative, as well as the ability to cooperate. In addition, the results of the Convention National Examination (UN) in 2013 organized by Kemdikbud decided that the determination of graduation to increase the credibility and reliability, then forward Natinal Examination measuring higher cognitive domain *(higher order thinking)*. Training the students to skilled, can be done by teachers practice the test that characterized HOTS. The problem faced by teachers is the ability in developing HOTS assessment instruments are still lacking, in addition the unavailability of assessment instrument designed

specifically to practice HOTS or higher order thinking skills of students. This is accordance with the results of Thompson research (2008, p.96) stated that the interpretation of Mathematics teacher from 32 people have difficulty interpreting skills of thinking in Bloom's Taxonomy and make test items for higher order thinking.

The problem, which occurs at schools, the test tend to be more testing the memory aspect is less to practice higher order thinking skills of students, the ability to think scientifically considered Indonesian children is still low as seen from the TIMSS survey results, one contributing factor, among others, students in Indonesia are less practiced in solving problems which measure HOTS, and the problems faced by teachers is the ability in the HOTS assessment instrument is still lacking and the unavailability of the assessment instrument designed specifically to practice HOTS, so it is necessary to develop HOTS assessment instruments. The development of higher order thinking skills students will generate: students proficiency in problem solving strategies to become good, confidence level students in Mathematics increased, and learning achievement of students in non-routine problems that require increasing higher order thinking skills (Butkowski, et al., 1994).

Form of assessment instrument consist of test and non-test instruments. Form of assessment instruments developed in this study using a multiple choice test instrument and descriptions. Multiple choice and essay test can be used for measuring HOTS or higher order thinking skills, it is in accordance with the opinion Brookhart (2010, p.33), Nitko & Brookhart (2011, p.223), Kubiszyn & Borich (2013, p.143), and Sumarna Surapranata (2007, p.137). The recommended approach to measure higher level thinking that by using the context-dependent item sets or a set of items which consist of an introduction and followed by choice answers and context-dependent item sets or exercises in interpreting. The introductory object to make HOTS items test, among other, using pictures, graphs, tables and so on are demanding students on the level application of the taxonomy of educational objectives and involve cognitive processes higher levels.

Based on the above issues, the HOTS assessment instruments need to be developed HOTS test items in multiple choice and essay in mathematics junior high class VIII first semester. HOTS assessment instruments developed aims to produce a valid and reliable instrument for measuring HOTS students. This research has its benefits, such as: the assessment instruments are valid and reliable can be used to measure HOTS students, as a reference to develop HOTS assessment instruments on the other Basic Competency (KD), and can be used by students as a practice in training HOTS.

**Research Methods**

This research is development research. The products that developed are HOTS assessment instruments in multiple choice test and essay test items. In getting a prototype development, this research was done on the adaptation of the Borg & Gall's model development. The 10 steps Borg & Gall's model development were adapted into seven developmental steps: (1) research and collecting information, (2) planning, (3) the initial product development, (4) limited testing, (5) the revision of initial product, (6) field testing, and (7) the revision of final product. The research and collecting information are carried out to study the concept based on the studies of relevant theory. The validation of assessment instrument is carried out to evaluate the validity of assessment instrument in the HOTS test items form. The validation is performed in the early stages of product development by three experts on mathematics education. The empirical test of HOTS test item is done by conducting a limited and field testing. Limited testing was conducted to 31 students of SMP Negeri 2 Banjarnegara. Field testing was conducted to 178 students of SMP Negeri 1 Banjarnegara, SMP Negeri 2 Banjarnegara, and SMP Negeri 2 Mandiraja. The data analysis of limited and field test is using classical test theory parameters to determine the quality of a HOTS test item empirically as a basic for revision and assembly of HOTS test item.

**Data, Instruments, and Data Collection Techniques**

The data in this study includes quantitative and qualitative data. These data aimed to give description on the quality of products that being developed. The qualitative data obtained from the results of the initial product expert validation HOTS test item, while quantitative data obtained from a HOTS test item product testing. The research instrument that developed in this research classified into two types, each of which is used to meet the criteria of valid and reliability.

Instrument for measuring the validity used validation sheet (review test questions) were analyzed qualitatively. Validation review from three aspects: material, construction, and language. The questions were considered valid or worthy based on the validator's assessment. Instrument for measuring the reliability used two sets of test questions, consists of a multiple choice questions and problem description. HOTS test item is examined individually and the results were analyzed quantitatively to know the estimated coefficient of reliability assessment instruments developed.

HOTS test item organized by HOTS indicators and Basic Competency (KD) indicators. HOTS indicator synthesized from indicators of critical and creative thinking by Nitko & Brookhart (2011, p.232), Arends & Kilcher (2010, pp.214-233), Presseisen (1985, p.45), Szetela (1993, p.143 ), Krulik & Rudnick (1999, p.139), O'Daffer & Thornquist (1993, p.40), Maite & Laura (2011, p.609), and Perkins (1985, p.58). The indicator is meant, among others: (1) identify and associate the relevant information from a situation/ problem, (2) make the right conclusions based on the information of a situation/ problem, (3) find the consistency/ inconsistency in an operations/ products, (4) assess an operation/ relevant products based on criteria/ standards, (5) blends the ideas/ strategies to solve a problem, (6) using the ideas/ the right strategies to solve a problem, (7) develop or create new alternative in resolving a problem.

Data collection techniques used by the researcher are as follows: (1) drafting the instruments that will be used in research, such as HOTS test item, scoring and assessment, (2) determine the validity of the content of the instrument with expert judgement or ask some mathematics education experts to validate the instrument that have been made, (3) do the revision of instruments complies with the suggestions of the validator, (4) testing research instruments, (5) determine the reliability, difficulty level, and distinguishing items, (6) do the revision instrument based on the analysis of the testing results.

**Data Analysis Techniques**

*Qualitative Analysis of HOTS Test Item*

Qualitative analysis of HOTS test item obtained from the validation sheet (test question study) that conducted in a descriptive qualitative. The data is the value of each test item numbers assessment results by the experts that analyzed by using the Aiken's V formula for calculating the content validity coefficient. Range of numbers V which can be obtained between 0 and 1.00.

*Quantitative Analysis HOTS Test Item*

The data obtained from the students response answers were analyzed by using the assistance software MicroCAT ITEMAN 3.00 for analysis multiple choice items, whereas the Microsoft Excel program assists the analysis of description item. The question analysis is used to determine the characteristics of items including difficulty level, differentiator, and deployment of answer choices/ options (distractor) for multiple choice items, while statistics item will be obtained characteristics device that is the average, standard deviation, difficulty level, differentiator, reliability coefficient, and SEM.

**Research Results and Discussion**

Development results

The developmental results in this research are valid and reliable HOTS assessment instrument in multiple choice test and essay test items of mathematics junior high class VIII first

semester. Developed assessment instruments has passed through two stages of assessment. The first stage of assessment was carried out to assess the validity of the assessment instrument conducted by the Mathematics education experts. The second phase assessment conducted field testing involving 178 students from three schools, the assessment focused on the characteristics of HOTS test items.

The process done in this development include the preparation of HOTS test items. HOTS test items designed assess by the validator expert, do revision to obtained the initial product of HOTS test items that ready to used as limited materials testing. The results of testing are limited, as a revision to the main product of HOTS test items that ready to used as a field testing. Estimation of reliability coefficient is obtained, criteria of difficulty level, differentiator, and alternative distractor the results of field testing, obtained the final product of HOTS test items that ready to used.

Product Testing Results

Validation by experts conducted to see the contents of the initial product. This validation aims to get input, suggestions for improvements, as well as an assessment of the initial product before conducted testing limited. The validation activities are carried out by providing initial product text in the form of lattice items and HOTS test items and validation sheet to three expert validator. Further assessment analysis of HOTS test items carried out in accordance with the assessment validator using the Aiken's V formula to calculate content validity coefficient. Data validation expert analysis result can be seen in Table 1 and Table 2 below.

Table 1. Analysis Validation of HOTS Test Item in Multiple Choice

| Test Item Number | Aiken's V Coefficient | Criteria |
|---|---|---|
| 1 – 30 | 0,67 – 1,00 | Eligible |

Table 2. Analysis Validation of HOTS Item Test in Essay

| Test Item Number | Aiken's V Coefficient | Criteria |
|---|---|---|
| 1 – 5 | 0,67 – 1,00 | Eligible |

Based on the results of analysis using the Aiken's V formula HOTS test item consisting of 30 multiple choice items and 5 description items, all stated feasibility. Nevertheless, there are some items that are fixed in according to the input and suggestions from the three validators that is about stem improvement in the formulation of sentence, the introductory material completeness on the stem, and lack of indicators according to the test item.

Limited testing results obtained information the time it takes to complete the HOTS test items, for multiple choice and essay, takes time each of the approximately 120 minutes. In addition, through the interpretation of the analysis items can know the quality of items based on the characteristics of the items that include the level of difficulty, differentiator, and also the spread of the answer choices/options (distractor) for multiple choice test and can also be known because statistics.

*Characteristics of HOTS Test Item in Multiple Choice Limited Testing Results*

Difficulty level of test item in multiple choice can be seen in Table 3 below.

Table 3. Difficulty Level of Initial Products HOTS Test Item in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| TK < 0,25 (Hard) | 4, 11, 15, 22, 23, 28, 29 | 7 | 23.33 |
| 0,25 ≤ TK ≤ 0,80 (Medium) | 1, 2, 3,5, 6, 7, 8, 9, 10, 12, 13, 14, 16, 17, 18, 19, 20, 21, 24, 25, 26, 27, 30 | 23 | 76.67 |
| TK > 0,80 (Easy) | - | 0 | 0 |

Based on Table 3 it can be known that the difficulty level ranging in the category as many as 23 items (76.67%).

Distinguishing items known by looking at the correlation coefficient point Biser (rpbis). In general distinguishing in multiple choice items can be seen in Table 4 below.

Table 4. Initial Product Distinguishing HOTS Test Items in Multiple Choice

| Kategori | Test Item Number | Sum | % |
|---|---|---|---|
| DP ≥ 0,40 (Good) | - | 0 | 0 |
| 0,30 ≤ DP ≤ 0,39 (Received without revised) | 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 | 25 | 83.33 |
| 0,20 ≤ DP ≤ 0,29 (Received with revised) | 1, 5, 6, 16, 30 | 5 | 16.67 |
| DP ≤ 0,19 (Denied) | - | 0 | 0 |

Based on Table 4 it can be known that the distinguishing range in received without revised categories ranges as many as 25 items (83.33%).

Deployment of answer choices / options (distractor) test item in multiple choice can be seen in Table 5 below.

Table 5. Effectiveness of Initial Product Distractor HOTS Test Item in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| $r_{pbis}$ positive key answer, Response ≥ 5%, and $r_{pbis}$ negative distractor | 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18,19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 | 25 | 83.33 |
| $r_{pbis}$ negative key answer, Response < 5%, and $r_{pbis}$ positive distractor | 1, 5, 6, 16, 30 | 5 | 16.67 |

Based on Table 5 it can be known that the item with spread of the answer choices/options (distractor) which serves both as many as 25 items (83.33%).

The results of item characteristics analysis above, the conclusion that the good and received without revised item, received with revised, and denied can be seen in Table 6 below.

Table 6. Result of Initial Product Characteristics Analysis HOTS Test Item in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| Received witout Revised | 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 | 25 | 83.33 |
| Received with Revised | 1, 5, 6, 16, 30 | 5 | 16.67 |
| Denied | - | 0 | 0 |

Based on Table 6 items were categorized as good and received without revised as many as 25 items (83.33%). Items were categorized as good and received without revised directly used in the main product. Items received with the revised categories by 5 items (16.67%), revised before being used in main product. Items were categorized as good and received without revised and revised reassembled into main product HOTS test items in multiple choice that will be tested in field testing.

*Statistics of HOTS Test Items in Multiple Choice Limited Testing Result*

Statistics of the initial product HOTS test items in multiple choice based on limited testing can be seen in Table 7 below.

Table 7. Statistics of Initial Product Analysis Result HOTS Test Items in Multiple Choice

| Statistic Scale | | | |
|---|---|---|---|
| Mean | 11.161 | SEM | 2.390 |
| Standard Deviation | 4.378 | Mean P. | 0.372 |
| Median | 10.000 | Mean Item-Tot | 0.327 |
| Reliability Coefficient | 0.702 | | |

Based on Table 7 this test reliability coefficients 0.702 and SEM 2.390. Mean P/difficulty level average of test items is 0.372, it means the items on this test is medium. Mean Item-Tot./

distinguishing average by calculating the average value of this test point biserial 0.327, it means the items on this test is good (received) indicates that the HOTS test items in multiple choice able to distinguish between groups of students above and under group.

*Characteristics HOTS Test Items in Essay Limited Testing Results*

Difficulty level of test items in essay can be seen in Table 8 below.

Table 8. Difficulty Level of Initial Product HOTS Test Items in Essay

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| TK < 0,25 (Hard) | 1e, 2b, 2c, 3d, 3e, 4a, 4b, 4c | 8 | 42.10 |
| 0,25 ≤ TK ≤ 0,80 (Medium) | 1a, 1b, 1c, 1d, 2a, 3a, 3b, 3c, 5c | 9 | 47.37 |
| TK > 0,80 (Easy) | 5a, 5b | 2 | 10,53 |

Based on Table 8 it can be known that the difficulty level ranges from medium category of 9 items (47.37%) and hard category of 8 items (42.10%).

Distinguishing test item in essay can be seen in Table 9 below.

Table 9. Distinguishing Initial Product HOTS Test Item in Essay

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| DP ≥ 0,40 (Good) | 1a, 1b, 1c, 1d, 3a, 3c, 4a, 4b, 5b, 5c | 10 | 52.63 |
| 0,30 ≤ DP ≤ 0,39 (Received without Revised) | 2a, 3b, 5a | 3 | 15.79 |
| 0,20 ≤ DP ≤ 0,29 (Received with Revised) | 1e, 2b, 2c, 3d, 3e, 4c | 6 | 31.58 |
| DP ≤ 0,19 (Denied) | - | 0 | 0 |

Based on Table 9 it can be known that the distinguishing range in good and received without revised and received without revised category of 13 items (68.42%) and received with revised category of 6 items (31.58%).

The results of item characteristics analysis above, the conclusion that the number of good and received without revised items, received with revised, and denied can be seen in Table 10 below.

Table 10. Results of Initial Product Characteristics HOTS Test Items in Essay

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| Good, Received without revised | 1a, 1b, 1c, 1d, 2a, 3a, 3b, 3c, 4a, 4b, 5a, 5b, 5c | 13 | 68.42 |
| Received with revised | 1e, 2b, 2c, 3d, 3e, 4c | 6 | 31.58 |
| Denied | - | 0 | 0 |

Based on Table 10 good and received without revised of 13 items (68.42%). Items were categorized as good and received without revised directly used in the main product. Items were catagorized received with revised of 6 items (31.58%), revised before used in main product. Good and received without revised items and which have been reassembled into the main product HOTS test items in essay that will be tested in field testing.

*Statistical of HOTS Test Items in Essay Limited Testing Result*

Statistical of the initial product HOTS test items in essay based on limited testing can be seen in Table 11 below.

Table 11. Statistical Analysis Results Preliminary Product Problem Description Test HOTS

| Statistic Scale | | | |
|---|---|---|---|
| Mean | 33.935 | SEM | 5.838 |
| Standard Deviation | 19.477 | Mean P. | 0.378 |
| Median | 30.000 | Mean Item-Tot | 0.493 |
| Reliability Coefficient | 0.910 | | |

Based on Table 11, reliability coefficient is 0.910 and SEM is 5.838 of this test. The difficulty level average is 0.378, it means that the items on this test is medium and distinguishing average 0.493 it means the item on this test is good (received). Distinguishing already good (received) indicates that the HOTS test item in essay able to distinguish between the students upper and lower groups.

Field testing conducted to determine the quality of HOTS test items based on the items characteristics and statistical of HOTS test items in multiple choice were developed from the results of limited testing.

*Characteristics of HOTS Test Items in Multiple Choice Results of Field Testing*

Difficulty level of multiple choice items can be seen in Table 12 below.

Table 12. Difficulty Level of Main Products HOTS Test Item in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| TK < 0,25 (Hard) | - | 0 | 0 |
| 0,25 ≤ TK ≤ 0,80 (Medium) | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 | 30 | 100 |
| TK > 0,80 (Easy) | - | 0 | 0 |

Based on Table 12 it can be seen that the difficulty level of HOTS test items in multiple choice in middle category.

Distinguishing items known by looking at the correlation coefficient point Biser (rpbis). In general distinguishing items in multiple choice can be seen in Table 13 below.

Table 13. Distinguishing Main Products of HOTS Test Items in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| DP ≥ 0,40 (Good) | 8, 15, 21, 26, 30 | 5 | 16.67 |
| 0,30 ≤ DP ≤ 0,39 (Received without revised) | 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14, 17, 19, 22, 23, 24, 27, 28, 29 | 19 | 63.33 |
| 0,20 ≤ DP ≤ 0,29 (Received with revised) | 1, 12 | 2 | 6.67 |
| DP ≤ 0,19 (Denied) | 16, 18, 20, 25 | 4 | 13.33 |

Based on Table 13 it can be seen that the distinguishing in good and received without revised category as many as 24 items (80%) and received with revised and replaced category as many as 6 items (20%).

Deployment of answer choices / options (distractor) test items in multiple choice can be seen in Table 14 below.

Table 14. Effectiveness Distractor of HOTS Test Items in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| $r_{pbis}$ positive key answer, Response ≥ 5%, and $r_{pbis}$ negative distractor | 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 17, 19, 21, 22, 23, 24, 26, 27, 28, 29, 30 | 25 | 83.33 |
| $r_{pbis}$ negative key answer, Response < 5%, and $r_{pbis}$ positive distractor | 1, 16, 18, 20, 25 | 5 | 16.67 |

Based on Table 14, it can be seen that the item with the spread of the answer choices / options (distractor) which serves both as many as 25 items (83.33%).

The results of the analysis of the characteristics of the above items, the conclusion that the number of good items and received without revised, received with revised, and denied can be seen in Table 15 below

Table 15. Results Analysis of Main Product Characteristics HOTS Test Items in Multiple Choice

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| Good and received without revised | 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 17, 19, 21, 22, 23, 24, 26, 27, 28, 29, 30 | 24 | 80 |
| Received with revised | 1, 12 | 2 | 6.67 |
| Denied | 16, 18, 20, 25 | 4 | 13.33 |

Based on Table 15 items were categorized as good and received without revised as many as 24 items (80%). Items were categorized as good and received without revised is used directly as the final product. Items received with the revised categories and replaced as many as 6 items (20%) are not used. Items were categorized as good and received without revised reassembled into final product of HOTS test items in multiple choice are ready for used.

*Statistical of HOTS Test Items in Multiple Choice Field Testing Results*

Statistical of main product HOTS test items based on field testing can be seen in Table 16 below.

Table 16. Statistical Analysis Results of Main Products HOTS Test Items in Multiple Choice.

| Statistic Scale | | | |
|---|---|---|---|
| Mean | 12.185 | SEM | 2,480 |
| Standard Deviation | 4.627 | Mean P. | 0.406 |
| Median | 11.000 | Mean Item-Tot | 0.330 |
| Reliability Coefficient | 0.713 | | |

Based on Table 16 Mean Item-Tot. / distinguishing average by calculating the average value of this test point biserial 0.330, it means the items on this test is good (received) indicates that the Hots test items in multiple choice able to distinguish students upper and lower groups. Mean P / difficulty level average of the matter is 0.406, it means the difficulty level of main product

HOTS test items in medium category. The reliability coefficient of this item is 0.713 and SEM is 2.480.

*Characteristics of HOTS Test Items in Essay Field Testing Results*

The difficulty level of item in essay can be seen in Table 17 below.

Table 17. Difficulty Level of Main Product HOTS Test Items in Essay.

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| TK < 0,25 (Hard) | 1d, 1e, 2b, 2c, 3d, 3e, 4a, 4b, 4c | 9 | 47.37 |
| 0,25 ≤ TK ≤ 0,80 (Medium) | 1a, 1b, 1c, 2a, 3a, 3b, 3c, 5b,5c | 9 | 47.37 |
| TK > 0,80 (Easy) | 5a | 1 | 5.26 |

Based on Table 17, it can be known that the difficulty level ranges in hard category as many as 9 items (47.37%) and medium category as many as 9 items (47.37%).

Distinguishing item description can be seen in Table 18 below.

Table 18. Distinguishing Main Product HOTS Test Items in Essay

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| DP ≥ 0,40 (Good) | 1b, 1c, 1d, 2a, 3b, 3c, 4a, 5b, 5c | 9 | 47.37 |
| 0,30 ≤ DP ≤ 0,39 (Received without revised) | 1a, 1e, 2b, 2c, 3a, 3d, 3e, 4b, 4c, 5a | 10 | 52.63 |
| 0,20 ≤ DP ≤ 0,29 (Received with revised) | - | 0 | 0 |
| DP ≤ 0,19 (Denied) | - | 0 | 0 |

Based on Table 18, it can be known that the distinguishing items in good and received without revised category, it indicates that the HOTS test items in essay able to distinguish between students upper and lower groups.

Analysis results of characteristics items above, the conclusion that the number of items which good and received without revised, received with revised, and denied can be seen in Table 19 below.

Table 19. Analysis Results of Characteristics Main Product HOTS Test Item in Essay

| Category | Test Item Number | Sum | % |
|---|---|---|---|
| Good and received without revised | 1a, 1b, 1c, 1d, 1e, 2a, 2b, 2c, 3a, 3b, 3c, 3d, 3e, 4a, 4b, 4c, 5a, 5b, 5c | 19 | 100 |
| Received with revised | - | 0 | 0 |
| Denied | - | 0 | 0 |

Based on Table 19, all items main product HOTS test items in essay good and received without revised category, it means the final product HOTS test items in essay already for used.

*Statistical of HOTS Test Items in Essay Field Testing Results*

Statistical of main product HOTS test items in essay based on field testing can be seen in Table 20 below.

Table 20. Statistical of Analysis Results Main Product HOTS Test Item in Essay.

| Statistic Scale | | | |
|---|---|---|---|
| Mean | 31.657 | SEM | 5.927 |
| Standard Deviation | 20.926 | Mean P. | 0.373 |
| Median | 3.000 | Mean Item-Tot | 0.508 |
| Reliability Coefficient | 0.920 | | |

Based on Table 20 this test reliability coefficients 0.920 and SEM for this test 5.927. The difficulty level average of item 0.373, it means the items on this test is medium and distinguishing average of items 0.508 it means items on this test is good (received). Distinguishing already good (received) indicates that the HOTS test items in essay able to distinguish between students upper and lower groups.

Product revision

Product revision were conducted to obtain a final product that meets the criteria of valid and reliable. Revision were conducted based on the results of assessment and analysis of assessment instruments at each stage of product testing. Product revision in this study consists of: the revised product of validation results, revised product limited testing results, and revised product field testing results.

Based on the results of evaluation after the expert assessment, limited testing, and field testing, assessment instruments developed undergone several revisions. First, the revised items based input and advice of validator. In general, the inputs and suggestions regarding improvements on stem like the formulation sentences, completeness introductory on the items, and the indicators are not in accordance with items. Second, the revised items based on limited testing initial product of HOTS test items. Items received with the revised category, revised based on the results of characteristics analysis of the items. Mostly conducted in spread of the less work answer choices/options (distractor) for HOTS test items in multiple choice and improvements in formulation of the sentence and completeness of the introductory information for the HOTS test items in essay. Third, revised items based on limited testing main products of HOTS test items. Item which received with revised and replaced category not used (discarded). Item which good and received without revised return to verified with HOTS indicator to know all indicators are represented. The verification results reassembled into final product of HOTS test items that ready for used.

Study of Final Products

The final product of this development research is the HOTS assessment instruments mathematics junior high class VIII first semester in HOTS test item. Based on the results of expert validation, limited testing, field testing, and improvements, as well as the data analysis can be known that the HOTS test items were developed has met the criteria of valid and reliable, as well as good quality items.

The validity of assessment instrument in the form HOTS test items based on validation criteria product development results that have been established. Validation is done by three experts from lecturer of postgraduate UNY. Validation by experts on HOTS test item products already meet the logical validity. Three validators state that HOTS test item products in multiple choice and essay are developed meet the criteria appropriate to used.

Assessment instrument reliability in form HOTS test items based on the analysis results of the main product HOTS test items. Reliability coefficient obtained from the analysis of HOTS test items in multiple choice is 0.713 with SEM 2.480, while the reliability coefficient of HOTS test items in essay is 0.920 with SEM 5.927.

The quality of assessment instruments in form HOTS test items based on the analysis results of the main product HOTS test items that analyze all items based on empirical data. HOTS test items in multiple choice have difficulty level average 0.406 (medium) and

distinguishing average 0.330 (good), and all distractors work well. While HOTS test items in essay have difficulty level average 0.373 (medium) and distinguishing average 0.508 (good).

**Conclusion and Suggestion**

Conclusion

Based on the results of research and discussion be concluded as follows: (1) the final product in this study resulted HOTS assessment instrument for measuring higher order thinking skills of junior high students class VIII. Assessment instrument in form HOTS test items which consists of 24 test items in multiple choice with four answer options and 19 test items in essay. The validity of instrument is evidenced by the results of expert assessment indicates that the instrument appropriate for used based on review of object aspects, construction, and language. The instrument also has met the criteria for reliable. (2) Test items in multiple choice have medium difficulty level, good distinguishing, all distractors work well, and test items in essay have medium difficulty level with good distinguishing.

Suggestion

Based on the research results and conclusions above, there are some suggestions final product utilization HOTS assessment instruments are as follows: (1) students can use the final product of HOTS assessment instruments as training object to practice higher order thinking skills, (2) junior high Mathematics teachers can use the final product of HOTS assessment instruments for measuring mastery of knowledge and higher order thinking skills of students, (3) final product of HOTS assessment instrument can be used as a reference in developing other Standard and Basic Competencies of HOTS assessment instruments.

**References**

Arends, R. I., & Kilcher, A. (2010*). Teaching for student learning becoming an accomplished teacher*. New York and London: Routledge Taylor and Francis Group.

Badan Standar Nasional Pendidikan (BSNP). (2006). *Standar isi untuk satuan pendidikan dasar dan menengah. Standar kompetensi dan kompetensi dasar*. Jakarta: BSNP.

Borg, W. R. & Gall, M.D. (1983). *Educational researcher: An introduction, (4th ed.)*. New York: Longman.

Brookhart, S. M. (2010). *How to assess higher order thinking skills in your classroom*. Virginia USA : SCD Alexandria.

Butkowski, J., Corrigan, C., Nemeth, T., & Spencer, L. (1994). Improving student higher order thinking skills in mathematics. *Theses, Mathematics Education Research*. Saint Xavier University-IRI, Field-Based Master's Program.

Conklin, W. (2012). *Higher-order thinking skills to develop 21st century learners*. Huntington Beach: Shell Educational Publishing, Inc.

Henningsen, M., & Stein, M.K. (1997). Mathematical task and student cognition: classroom based factors that support and inhibit level mathematical thinking and resaoning. *Journal for research in mathematics education, Vol. 28 No. 5. (Nov., 1997), pp. 524-549.*

Krulik, S., & Rudnick, J. A. (1999). Innovative tasks to improve critical and creative thinking skills. Dalam Lee V. Stiff & Frances R. Curcio (Editor), *Developing mathematical reasoning in grades K-12,1999 yearbook.* Reston, VA: The National Council of Teachers of Mathematics, Inc.

Kubiszyn, T. & Borich, G. D. (2013). *Educational testing & measurement. Classroom application and practice, (10th ed.).* New York: John Wiley & Sons, Inc.

Maite, G & Laura, B. (2011). Effect of a play program on creative thinking of preschool children. *Journal of Psychology, vol. 14, num. 2, 2011, pp. 608-618,* Universidad Complutense de Madrid. Espana, Diambil pada tanggal 08 Oktober 2013, dari http://www. redalyc.org/articulo.oa? id=17220620009

Mullis, I. V. S., Martin M. O., Foy P., & Arora A. (2012). *TIMSS 2011 international results in mathematics.* Boston: TIMSS & PIRLS International Study Center.

NCTM. (2000). *Principles and standards for school mathematics.* Reston, VA: The National Council of Teachers of Mathematics, Inc.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of student, (6th ed.).* Boston: Pearson Education.

O'Daffer, P. G., & Thornquist, B. A. (1993). Critical thinking, mathematical reasoning, and proof. Dalam P. S. Wilson (Editor), *In research ideas for the classroom: High school mathematics* (pp. 39-56). New York: Maxwell Macmillan International.

Perkins, D. N. (1985). What Creative Thinking Is. Dalam Arthur L. Costa (Edited), *Developing minds: A resource book for teaching thinking* (pp. 43-48). Alexandria, Virginia: ASCD.

Presseisen, B. Z. (1985). Thinking skills: meanings and models. Dalam Arthur L. Costa (Edited), *Developing minds: A resource book for teaching thinking* (pp. 43-48). Alexandria, Virginia: ASCD.

Resnick, L. B. (1987). *Education and learning to think.* Washington, D.C: National Academy Press.

Sumarna Surapranata. (2007). *Panduan penulisan tes tertulis. Implementasi kurikulum 2004.* Bandung: PT Remaja Rosdakarya.

Szetela, W. (1993). Facilitating communication for assessing critical thinking in problem solving. Dalam Webb, N. L. & Coxford, A. (Editor), *Assessment in the mathematics classroom, 1993 yearbook.* Reston, VA: The National Council of Teachers of Mathematics, Inc.

Thompson, T. (2008). Mathematics teachers' interpretation of higher-order .hinking in bloom's taxonomy. *International Electronic Journal of Mathematics Education:* Diambil pada tanggal 25 April 2013, dari http://www.doaj.org.

Van de Walle, J. A. (2007). *Elementary and middle schoolmathematics: teaching developmentally, (6th ed.).* United States of America: Pearson Education, Inc.