USEJ 9 (1) (2020)

Unnes Science Education Journal

http://journal.unnes.ac.id/sju/index.php/usej

THE DEVELOPMENT OF TWO-TIER MULTIPLE CHOICE ASSESSMENT INSTRUMENT TO MEASURE HIGHER ORDER THINKING SKILLS (HOTS) OF THE STUDENTS ON EXCRETION SYSTEM MATERIAL

Yunita[1], N. R. Dewi[1] ✉

[1]Science Education Study Program, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang (UNNES), Indonesia

Article Info

Article History:
Received April 2019
Accepted July 2019
Published Februari 2020

Keywords:
assessment, two-tier multiple choice, higher order thinking skills

Abstract

Science learning based on 2013 curriculum requires the students to be able at analyzing, evaluating, and creating to achieve higher order thinking skills (HOTS). The results of observation at SMP N 2 Sukolilo showed that there was no task or question which was specifically used to measure HOTS of the students. The aim of this study is to analyze the characteristics of two-tier multiple choice assessment instrument and HOTS ability of the students. This research was conducted by using Research and Development (R&D) method. Samples were taken by random sampling technique from the population of class IX students of SMP N 2 Sukolilo in the academic year 2019/2020, then class IX C was selected for small-scale trial, class IX B was selected for large-scale trial, and class IX A was selected for trial use. Data was taken by some methods such as test, questionnaire, and documentation. The characteristics of two-tier multiple choice assessment instrument was analyzed by expert judgment and empirical testing. Whereas, HOTS level of the students was analyzed using the HOTS Category Scale. The results of the analysis of the characteristics of two-tier multiple choice assessment instrument indicated that the developed instrument had fulfilled the requirement as a good instrument, namely (1) valid, with the result of expert validation as many as 79% (feasible) and the results of the item analysis was 0.49 (sufficient); (2) reliable, with the reliability result was 0.40 (reliable); (3) relevant, representative, practical, and specific as indicated by the results of the teacher's questionnaire responses with the percentage 75% (good); (4) discriminatory power, with the average of analysis result was 0.36 (good); and (5) proportional / difficulty level, with the analysis result of most questions in the moderate category. Furthermore, the analysis results of HOTS level of the students showed that the students on the large-scale trial were mostly in the moderate category (67%) and some of them in a small portion were in the low category (33%). While, HOTS level of the students on the trial use showed that most of the students were in the moderate category (50%), some of them were in the low category (30%), and a small portion of them were in the high category (20%). The conclusion of this study is the developed two-tier multiple choice assessment instrument has been feasible, because it has met the requirements as a good instrument and can be used to measure HOTS of the students.

p-ISSN 2252-6617
e-ISSN 2252-6232

✉Corresponding author:
N. R. Dewi
Science Education Study Program, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang (UNNES), Indonesia
E-mail: noviratnadewi@mail.unnes.ac.id

INTRODUCTION

The implementation of primary and secondary education, as stated in Government Regulation Number 17 of 2010 concerning Management and Implementation of Education aims to build a foundation for developing potential students (Dewi, 2019). Natural Sciences is a subject that implements a scientific approach, thus it is expected that the learning carried out based on inquiry (Dewi, 2018). The students are asked to be directly involved and find their own knowledge about something, because the nature of science generally has three components, namely process, product, and scientific attitude. The scientific process includes observing, classifying, predicting, designing, and carrying out experiment, while the scientific product can be in the form of fact, principle, concept, law and theory, then the scientific attitude, namely curiosity, caution, objectivity, and honesty. The students are required to have skill in order to study the existing natural events in scientific ways to gain the knowledge.

Science learning is expected to provide knowledge (cognitive) which is the main goal of learning. Science education plays an important role in the development of scientific literacy, scientific methodology, scientific concept, and process skill (Dewi, 2019). Besides providing knowledge, learning science is also expected to provide skill (psychomotor), the ability of scientific attitude (affective), comprehension, habit, and appreciation as the purpose of education in general (Trianto, 2012).

Science learning based on 2013 curriculum aims to develop students' knowledge, understanding, and analytical skill towards the natural environment and surroundings. The students are required to be skilled at analyzing, evaluating, and creating. In the revision of Bloom's Taxonomy, these three skills are at the cognitive level C4, C5, and C6 which are developed to achieve Higher Order Thinking Skills. This Higher Order Thinking Skills begins to be developed in 2013 curriculum. The development of 2013 curriculum needs to be done because of the various challenges faced, both internal and external challenges. External challenges in education are facing the 21st century and globalization. 21st Century skills that students need to possess are creativity, critical thinking/problem solving, communication, and collaboration. There are two kinds of problem solving skills, namely academic problem and authentic problem. Academic problem is related to the knowledge being studied, while authentic problem is related to problem encountered in everyday life. For this reason, the learning process requires complex problem solving exercises. In these training activities, students need intelligent and creative thinking processes which are the characteristic of higher-order thinking.

The higher order thinking skills (complex) will make the students accustomed to facing difficult problems. The students who have higher order thinking skills will be able to compete in a global world. Learning higher-order thinking since the education unit will create extraordinary students who are capable of solving problems in the future. The higher order thinking skills on the students which equipped with good skills and mental attitudes is the ultimate goal of 2013 Curriculum.

The success of this ultimate goal can be known through assessment or evaluation conducted by the teacher. The teacher needs an appropriate assessment which then used to measure that ability, namely a test in the form of Higher Order Thinking Skills (HOTS) questions. Besides being an assessment tool, HOTS questions can help teacher to improve the quality of the questions. In addition, HOTS questions also train the students to work on national and international Olympic standard questions. Through HOTS questions, the students' curiosity and understanding of the material also increased.

The results of Program for International Student Assessment (PISA) survey in 2018 stated that Indonesia was ranked 10th lowest out of 79 countries in the category of mathematics, literacy, and science. In the context of educational theory, the condition of the students in Indonesian was still at the level of Lower Order Thinking Skill (LOTS) or low thinking skills. Moreover, the results of the latest TIMSS survey conducted in 2015 showed that the students in Indonesia had not shown satisfactory achievement. Indonesia scored 397 in the field of Science, and ranked Indonesia in 45th out of 48 countries.

The results of that survey showed that in average the students in Indonesia were more likely to master easy and moderate questions, which only required the students to have the ability to remember and understand. The questions tested in TIMSS were questions that trained the students' thinking skills not only in understanding aspect, but also in application and reasoning aspects (Balitbang, 2011).

Agustin (2013) states that the questions used to measure standard learning outcomes in Indonesia are in the form of summative test and the evaluation made by the educator almost never bring up the questions that are measure students' higher-order thinking skills. Furthermore, the questions that are tested tend to only measure the mastery of science product which only aims to find out how far the knowledge possessed by the students without training their thinking skills. This is in line with the results of observation made by the researcher towards the assessment instrument used in SMP N 2 Sukolilo. The results of observation shows that the questions used by the science teacher only measure the students' ability in remembering and understanding, there is no specific question used to measure students' HOTS.

The higher order thinking skills of the students can be developed and trained during the learning process. The higher order thinking skills that have been achieved by the students can be measured through learning evaluation in the form of appropriate assessment. Evaluation instrument to measure higher order thinking skills can use various types of assessment such as modified multiple choice, short answer construction, and long answer construction as has been done by Ramirez (2008). One of the alternatives of modified multiple choice that can be used to measure higher-order thinking skills is in the form of two-tier multiple choice question.

Haladyna & Downing (1989) confirm that the advantages of two-tier multiple choice questions namely can be used to measure students' cognitive ability at a higher level (Higher Order Thinking). The form of two-tier multiple choice question can be used to help the teacher in testing the students understanding and help the teacher in identifying misconceptions students might have. Cullinane (2011) states the inclusion of reasons at the second level in two-tier multiple choice questions can be used to improve higher-order thinking skills and see the students' ability in reasoning. The inclusion of reasons at the second level in this question can be used to reduce the occurrence of advantages and disadvantages that are often become the weaknesses of ordinary multiple choice question. The assessment of questions which are objective, easy, and fast are the advantages of two-tier multiple choice question compared to other higher-order thinking skills questions. Based on the description above, this study aims to develop two-tier multiple choice question assessment instrument to measure Higher Order Thinking Skills (HOTS) of the students on excretion system material ". According to Muhammad Yaumi (2013) an assessment instrument can be called as an assessment tool, in the form of material used to obtain facts using the chosen method. Zainal Arifn (2011) reveals that the characteristics of good evaluation instrument are "(1) valid, (2) reliable, (3) relevant, (4) representative, (5) practical, (6) discriminatory, (7) specific, and (8) proportional ".

Ernawati (2017) argues that higher order thinking skills (HOTS) is a way of thinking that no longer only memorizes verbally, but also interprets the nature of the values contained in it, to be able to interpret that meaning requires an integralistic way of thinking with analysis, synthesis, and association to draw the conclusion towards the creation of creative and productive ideas. Higher-order thinking skills are based on lower skills such as discrimination, simple application, and analysis, as well as cognitive strategy related to prior knowledge of the subject content. Generally, HOTS questions measure the ability on the some domains such as analyzing (C4), evaluating (C4), and creating (C6). One form of the test assessment techniques that can be used to measure students' higher order thinking skills (HOTS) is a two-tier multiple choice assessment form.

Amir (1994) found that the multiple choice question method is an effective and sensitive tool in learning assignments, by changing some things which become the limitations of the common multiple choice test. The result is that it is suggested that a multiple choice test be arranged that asks the students to explain in answering. The product appears in the modification of multiple choice test is two tier multiple choice diagnostic test, which is specifically developed to identify alternative conceptions in a limited and predetermined area.

According to Treagust (2006) the development of two-tier diagnostic test can be used as an effective way to measure students' concepts. The first tier of each item in the test is a proportional statement and part of the concept map made in the form of multiple choice. The second tier contains the reasons students must choose to explain the answers on the first tier in the form of multiple choice. Rusilowati (2015) revealed that through this way the teacher can know the students who answered right for the right reason and students who answered right for the wrong reason.

This study is used to analyze the characteristics of developed assessment instrument as well as to analyze the HOTS ability of the students which is measured by using a two-tier multiple choice assessment instrument. The development of a two-tier multiple choice assessment instrument is expected to be used as a reference for the teacher to make a test used to train the students in working on HOTS questions. In addition, it is expected that it can be an evaluation tool to train and measure the HOTS of the students which can be implemented in the application of examination.

METHODS

This study was conducted at SMP N 2 Sukolilo located in Pati City. This study was conducted in the odd semester in the academic year 2019/2020. The samples used in this study were 12 students of class IX C on a small scale trial, 18 students of class IX B on a large scale trial, and 20 students of class IX A on the trial use.

The researcher in this study used a type of Research and Development (R&D). Research and development is a process or steps to develop a new product or to complete an existing product and can be answered or it is accountable product (Pramana, 2014). This research was conducted using a research and development approach (Research and Development) using a model adapted from Sugiyono (2012) that had been modified (Setyanto, 2015), which consisted of 9 steps, namely: (1) potential and problems; (2) data collection; (3) product design; (4) design validation; (5) design revision; (6) product trial; (7) product revision; (8) trial use; (9) product revision.

Data collection techniques used in this study were in the form of test, questionnaire (validation from material expert and assessment expert), and documentation. The data analysis techniques used were qualitative analysis and quantitative analysis consisting of expert validation analysis, questionnaire response analysis, and item analysis.

RESULTS AND DISCUSSION

The Characteristics of Assessment Instrument

Valid

Validity is a measure that shows the levels of validity of an instrument. The valid criteria in this study are categorized into design validation by the expert as well as item validity.

Design validation in this study was carried out by two experts, namely assessment expert and material expert. Each expert consisted of 3 validators. The assessment instrument is said to be feasible if they obtain a percentage> 63%. Before being validated, the assessment and material expert provided some input. Each input from the validators was recapitulated into one, evaluated and then sought solutions for the improvement (Dewi, 2016).

The assessment expert provided some suggestions for the developed instrument because the whole questions did not include HOTS indicator, the form of the questions was not varied, and the writing system of the questions was not in the correct format. While the suggestions given by the material experts were because there were still several items which were not in accordance with the truth of the concept and there were still some unclear pictures. These suggestions were then used as material for the improvement of the developed instrument. The improvements made by the researcher, namely changing the form of the questions, adjusting HOTS indicator with the item, checking the truth of the concept of the questions, and changing the pictures on the questions. The improvement was made to meet the criteria by the assessment expert and the material expert so that the instrument could be said to be feasible.

The results of assessment by the assessment expert can be seen in Table 1.

Table 1. The Results of Assessment Expert Validation

Aspect	Score Percentage (%)		
	Validator I	Validator II	Validator III
Question clue	87,5	87,5	75
Component contents	75	75	75
Evaluation Component	92	75	71
Use of Language	75	75	75

Table 1 shows that the results of the assessment of two-tier multiple choice instrument that has been validated by the assessment experts obtained the average percentage as much as 76%.

The results of the assessment by the material expert can be seen in Table 2.

Table 2. The Results of Expert Material Validation

Aspect	Score Percentage (%)		
	Validator I	Validator II	Validator III
Relevance	75	100	75
Accuracy	75	90	75
Use of Language	75	100	75

Table 2 shows that the results of the assessment of two-tier multiple choice instrument that has been validated by the material experts obtained the average percentage as many as 82%.

The results of the assessment from each expert shown in Tables 1 and 2 show that the developed assessment instrument is feasible. This feasible criterion is given by the assessment expert because the questions developed are in accordance with (Core Competence) CC and (Basic Competence) BC on the excretion system material. This is in line with the statement of Permendikbud No. 104 of 2014 concerning Assessment of Learning Outcomes at the Primary and Secondary Education Level. "The completeness of mastering the substance is the completeness in learning Basic Competence (BC) which is the level of mastery of the students on certain BC at a minimum mastery level or above". Furthermore, this is also in line with the writing systems of multiple choice questions in the Ministry of National Education (2008), namely the questions must be in accordance with the indicators (means that the questions must ask for behavior and material to be measured in line with the formulation of indicators in the grid). The instrument developed by the researcher has met the feasible category by the assessment expert and the material expert so that it can be used for trial in the school.

The next analysis of validity of the instrument is done by analyzing the validity of the items. Item analysis is carried out with the aim of finding out the validity of each item tested to the students. The analysis of validity of these items is carried out in small-scale trial, large-scale trial and trial use. The results of the validity analysis of the items or questions can be seen in Table 3.

Table 3. The Results of Validity Analysis of The Items

Trial	Number of Item	Total Item	Criteria
Small-scale	1, 2, 3, 5, 6, 9, 11, 13, 15, 16, 17. 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 30	22	Valid
	4, 7, 8, 10, 12, 14, 21, 22, 31, 32	10	Invalid
Large-scale	1, 2, 3, 4, 5, 6, 7, 9, 13, 14, 15, 16, 18, 19, 20, 23, 24, 25, 28, 29, 30, 32, 11, 22	24	Valid (used)
	8, 10, 12, 17, 21, 26, 27, 31	8	Not valid (removed)
Usage	1, 2, 3, 4, 5, 6, 7,8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	24	Valid

The results of the calculation of the validity of the questions on a small-scale trial showed that of the 32 questions worked on by students, 22 were included in the valid criteria. The results of the calculation of the questions validity on a large-scale trial showed that of 32 questions worked on by students, 24 questions were included in the valid criteria. While the results of the calculation of the questions validity on the usage test show that of the 24 questions done by students, all of the questions are included in the valid criteria. The instrument is said to be valid if it has a validity coefficient (r_{xy}) around 0.50. The validity coefficient is the result of estimating the validity of a measurement which is empirically stated and usually expressed by the correlation between the distribution of test scores and the distribution of criteria scores. The validity coefficient only has meaning if it has a positive value. The closer the value to 1.00, the more valid the test result is. The interpretation of validity and reliability coefficients are both relative, in general, the estimation of validity about 0.50 and it can be considered satisfactory, whereas validity coefficients less than 0.30 are usually considered unsatisfactory (Azwar, 2011). Validity is absolutely necessary for a measurement tool or test instrument so that the measurement objectives are relevant to the data needed or obtained.

Reliability

The results of the test is said to have high reliability if it gives relatively fixed results when used on another occasion (Arikunto, 2009). The results of the reliability analysis of the questions can be seen in Table 4.

Table 4. Reliability Analysis of the Questions

Trial	r_{11} Value	r_{tabel} Value	Criteria
Small-scale	0,759	0,349	Reliable
Large-scale	0,755	0,349	Reliable
Usage	0,854	0,404	Reliable

The two-tier multiple choice instrument for measuring students' HOTS developed by the researcher is reliable. This can be seen from the scores obtained, the scores obtained by the students are relatively the same even they are given repeated measurement, namely in large-scale trial, small-scale trial, and usage trial. In addition, the results of the calculation of the reliability of the questions in small-scale trial, large-scale trial, and usage trial show that the questions are reliable, where the price of r_{11} > 0.60. This is in line with the statement of Sugiyono (2012): "An instrument is declared reliable, if the reliability coefficient is at least 0.60". Based on this opinion, it can be seen that an instrument is declared reliable if the value ≥ 0.60, while the instrument is declared unreliable if the value <0.60. Reliability is needed to know the extent to which the results of a measurement can be trusted and in order the questions developed can be used at other opportunities or in the future.

Relevant, Representative, Practical, and Specific

An instrument can be said (1) relevant if it is in line with the basic competence and indicator; (2) representative if the material used is in accordance with the learning; (3) practical if easy to use; and (4) specific if the instrument is specifically used for the object being measured. The analysis of the four characteristics carried out by giving a questionnaire response to the teacher. The results of teacher response analysis can be seen in Table 5.

Table 5. The Results of Teacher Response Questionnaire

Characteristics	Item Statement	Maximum Score	Total Score	Percentage (%)
Relevant	The development of science assessment instrument in accordance with CC and BC	8	6	75
	In the development of assessment instrument there is the grid questions		6	75
	The Directions for working on the questions are clearly conveyed		6	75
	The developed assessment instrument in accordance with the HOTS indicator and the grid questions given		6	75
Representative	Scientific terms used are easy to understand	8	6	75
	The questions used are appropriate to the level of student development		6	75
	The use of pictures and tables in the questions are clear		6	75
	The language used in the questions is easy to understand		6	75
	Indonesian used in accordance with official Indonesian and Malaysian spelling system		6	75
Specific	The developed assessment instrument make it easier for teacher to measure students' HOTS	8	6	75
	Can be used as an evaluation tool for excretion system material		6	75
Practical	Flexible when used by other teachers to measure students' HOTS		6	75

The developed instrument can be said to be relevant because it is in accordance with competency standard, basic competence, and indicator that have been set. In addition, the developed instrument is also in line with the domain of learning outcomes. This can be seen from the results of the assessment by the teacher which states that the developed instrument is consistent with the HOTS indicators as well as material indicators on excretion system. In addition, the developed instrument is in accordance with bloom's taxonomy (C4 to C6).

The results of data analysis of teacher response questionnaire show that the developed assessment instrument is good, with the acquisition of an average percentage is 75%, while the results of the questionnaire data analysis of student response obtained an average percentage 78% which is included in good category.

Discriminatory

The discriminatory power of question is the ability of an item to distinguish students who have high-ability from students who have low-ability (Arikunto, 2009). To find out whether a measuring instrument is sufficiently discriminatory or not, it is usually based on a discriminatory power test of the measuring instrument. The analysis results of discriminatory power of the questions can be seen in Table 7.

Table 7. The Analysis Results of Discriminatory Power of The Questions

Trial	Number of Item	Total Item	Criteria
Small-scale	6, 16, 24, 29, 30	5	Very good
	13, 14, 15, 20, 25	5	Good
	3, 9, 18, 19	4	Enough
	1, 2, 4, 5, 7, 8, 10, 11, 12, 17, 21, 22, 23, 26, 27, 28, 31, 32	18	Bad
Large-scale	2, 13, 15, 29	4	Very good
	6, 16, 18, 23, 25	5	Good
	1, 3, 4, 5, 7, 9, 11, 14, 19, 20, 22, 24, 28, 30, 32	15	Enough
	8, 10, 12, 17, 21, 26, 27, 31	8	Bad
Usage	4, 13, 14, 18, 20, 21, 22	7	Very good
	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 19, 23, 24	17	Good
	–	–	Enough
	–	–	Bad

Based on the results of the analysis results of discriminatory power of the questions in small-scale trial, the questions are mostly in the bad category. The analysis results of discriminatory power of the questions in large-scale trial indicate that most of the discriminatory power of the questions are in the Enough category. While the results of the analysis results of discriminatory power of the questions in the trial use are mostly in the good category (> 0.25%). Questions that have a good discriminatory power are included in good category in the classification of discriminatory power (Daryanto, 2012). Good items must have a discriminatory power index of at least 0.25% or even 0.35%. The items or questions which have discriminatory power less than 0.25% are considered to be not feasible (Oller, 1978). Discriminatory power analysis is carried out with the aim to find out the ability of the questions in the developed assessment to distinguish students who are smart (upper group) and students who are less (lower group).

Proportional

Proportional means that a measuring instrument must has a proportional difficulty level between difficult, moderate and easy. The results of the difficulty level analysis can be seen in Table 8.

Table 8. The Results of Difficulty Level Analysis

Trial	Number of Item	Total Item	Criteria
Small-scale	2, 3, 7, 8, 10, 12, 17, 23, 26, 27, 28, 31	12	Difficult
	1, 4, 5, 6, 11, 13, 15, 16, 18, 19, 20, 22, 25, 29, 30	15	Moderate
	9, 14, 21, 24, 32	5	Easy
Large-scale	1, 8, 9, 10, 19, 20, 21, 27, 31	9	Difficult
	2, 4, 7, 11, 12, 13, 14, 15, 16, 17, 18, 22, 23, 26, 28, 29, 30, 32	18	Moderate
	3, 5, 6, 24, 25	5	Easy
Usage	10, 15, 16, 23	4	Difficult
	4, 5, 8, 12, 13, 14, 17, 18, 20, 21, 22, 24	12	Moderate
	1, 2, 3, 6, 7, 9, 11, 19	8	Easy

The two-tier multiple choice instrument to measure HOTS of the students developed by the researcher has difficulty level which is included in good category. This can be seen from the results of difficulty level analysis of the questions in small-scale trial, large-scale trial, and usage trial. The results of results of difficulty level analysis of the questions in small-scale trial, large-scale trial, and usage trial show that from the whole questions, dominated by questions with Moderate difficulty level category. This is consistent with Arikunto (2009): "Judging from the difficulty level, too easy questions do not stimulate students to solve them, while too difficult questions can cause students to despair quickly. Therefore, a good question is a question that has a balanced level of difficulty which means that the question is not too easy and not too difficult with the difficulty index or question in the moderate criteria".

Higher Order Thinking Skills (HOTS) of the Students

Higher Order Thinking Skills (HOTS) is defined as the wider use of the mind to find new challenges. The advantages of two-tier multiple-choice test compared to common multiple-choice test are that it allows us to assess two aspects in one phenomenon (symptoms). At the first level the students are asked to answer symptoms that occur, then at the second level the students are asked to explain it. This allows us to assess the understanding of the students concepts (Tuysuz, 2009). In addition, the advantages of two-level multiple choice, the students are only considered correct if they answer both levels correctly, thereby reducing the level of assessment errors.

The developed assessment instrument in this study is two-tier multiple choice assessment instrument to measure HOTS of the students in learning science material "Excretion System". The assessment instrument in this study contains 32 multiple choice questions that are adjusted to HOTS indicator and material indicator for excretion system. After testing and analyzing the items or questions, 24 valid questions are obtained, so that the trial use in this study is in the form of a two-tier multiple choice assessment instrument which contains 24 multiple choice items.

HOTS level analysis of students is carried out on large-scale trial and trial use. Data obtained in a large-scale trial with 18 students, the results can be seen in Table 9.

Table 9. The Level of Higher Order Thinking Skills (HOTS) for Large Scale Trial

HOTS Category	Number of Students	Percentage (%)
Very Low	0	0
Low	6	33
Moderate	12	67
High	0	0
Very High	0	0

Table 9 shows that the HOTS level of the students in large-scale trial is mostly in the moderate category (67%) and a small proportion is in the low category (33%). The HOTS analysis of the students on the trial use can be seen in Table 10.

Table 10. The Level of Higher Order Thinking Skills (HOTS) for Trial Use

HOTS Category	Number of Students	Percentage (%)
Very Low	0	0
Low	6	30
Moderate	10	50
High	4	20
Very High	0	0

The results of large-scale trial indicate that most of the higher order thinking skills (HOTS) are in the moderate category (67%), and some are in the low category (33%). While the results of the trial use show that the higher order thinking skills (HOTS) are mostly in the moderate category (50%), some are in the low category (30%), and a small portion are in the high category (20%). The difference in the results from the two trial conducted by the researcher can be influenced by several factors.

Higher order thinking skill is essentially one form of learning outcomes which is influenced by various factors. High and low of this skill is influenced by model and learning media used by the teacher and the students' ability. According to Slameto (2010) the factors that influence learning outcomes are (1) Internal factor include: Physical factor consists of health factor and disability factor; Psychological factor consists of intelligence, attention, interest, talent, motive, maturity, and readiness. The fatigue factor is both physical fatigue and spiritual fatigue; (2) External factor include: Family factor consists of the way parents educate, relationship between family members, the atmosphere of the home, the situation of family economic, the understanding of parents, and cultural background. School factor consists of teaching method, curriculum, teacher-student relation, student-student relation, school discipline, instructional tool, school's time, standard lesson above the size, building condition, learning method, and homework or assignment. Community factor consists of student activities in the community, mass media, friend, and community life form. Therefore, the effort to develop higher-order thinking skills need to be done thoroughly by considering these factors.

Based on the results of research conducted by Zainuddin (2016) the increase in HOTS of the students can be obtained through physics experiment activity accompanied by the use of effective learning model. HOTS can increase in line with the experimental process that stimulates the students to think about designing a study, analyzing the results of an experiment and paying attention to the results of the final reflection. In addition, the research conducted by Budsankom (2015) shows that the classroom environment, psychological and intellectual characteristics of the students directly influence HOTS that is 96.8%. The classroom environment is one of the factors in influencing HOTS, it can be caused by a good environment that can help the students in facilitating the thought process.

Based on some of the factors above, the factors that influence the difference in the results of HOTS level of the students in large-scale trial and trial use include: (1) students who take the trial use are all female students, so that the classroom environment is more conducive than students on large-scale trial where in the class there are male students. The condition of the classroom environment can influence the concentration of the students during the learning; (2) students in the trial use have a higher level of intelligence (learning ability) than students in large-scale trial so that the ability to answer questions between students in these two classes is also different.

Two-tier multiple choice instrument to measure the HOTS of students that is developed by the researcher is feasible because it meets the characteristics of good instruments, namely: (1) valid, (2) reliable, (3) relevant, (4) representative, (5) practical , (6) discriminatory, (7) specific, and (8) proportional ", and can be used to measure higher-order thinking skills of the students.

CONCLUSION

The two-tier multiple choice instrument for measuring HOTS of the students developed by the researcher has been feasible based on expert judgment and empirical testing. The criteria are feasible because the developed instrument has met the requirements of a good characteristics of instrument, namely: valid, reliable, relevant, representative, practical, discriminatory, specific, and proportional. The analysis results of HOTS level of the students showed that the students on large scale trial are mostly in the moderate category (67%) and a small portion of them are in the low category (33%). While the HOTS level of the students on the trial use shows that most are in the moderate category (50%), some are in the low category (30%), and a small portion of them are in the high category (20%).

REFERENCES

Agustin, V. N. 2013. Peningkatan Aktivitas dan Hasil Belajar Siswa melalui Model Problem Based Learning (PBL). *Journal of Elementary Education. 2*(1) 36-44.

Amir, R. & Tamir. 1994. In Depth Analysis of Misconception as a Basis for Developing Research-Based Remedial Instruction: The Case of Photosynthesis. *The American Biology Teacher*. *56*(2) 94-100.

Arikunto, S. 2009. *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Azwar, S. 2011. *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar.

Balitbang. 2011. *Survei Internasional TIMSS (Trends In International Mathematics and Science Study)*. Online.

Budsankom, P. *et, all*. 2015. Factors Affecting Higher Order Thinking Skills of Students: A Meta-Analytic Structural Equation Modeling Study. *Educational Research and Reviews*: 2639-2652.

Cullinane., Alison, & Maeve, L. 2011. *Two-tier Multiple Choice Question: An Alternative Method of Formatif Assessment for First Year Undergraduate Biology Students*. Limerick: National Center for Excellence In Mathematics and Education Science Teaching and Learning (NCE-MSTL).

Daryanto. 2012. *Evaluasi Pendidikan*. Jakarta: Rineka Cipta.

Depdiknas. 2008. *Panduan Penulisan Butir Soal*. Jakarta: Direktorat Pembinaan Sekolah Menengah Atas.

Dewi, N. R., Isa, A. 2016. Pengembangan Perangkat Pembelajaran IPA Berbasis Pendidikan Multikultural Menggunakan Permainan untuk Mengembangkan Karakter Siswa. *Unnes Science Education Journal. 5*(1) 1098-1108.

Dewi, N. R., Isa, A., Fitria, N. A., & Muhammad, T. 2018. The Effect of Inquiry-Based Independent Worksheet Using ICT Towards Science Learning to Embody the Student's Creativity and Characters. *International Journal of Engineering & Technology. 7*(2.29) 574-580.

Dewi, N. R., Lailatul, M., Septia, N., & Ida, W. 2019. The Development of Contextual-Based Science Digital Storytelling Teaching Materials to Improve Students' Critical Thinking on Classification Theme. *Journal Of Turkish Science Education. 16*(3) 364-378.

Ernawati, L. 2017. Pengembangan High Order Thinking (HOT) Melalui Metode Pembelajaran Mind Banking Dalam Pendidikan Agama Islam. *Prosiding 1st International Conference on Islamic Civilization ans Society (ICICS).* Lamongan : Universitas Islam Darul 'Ulum.

Haladyna, T. M. & Downing, S. M. 1989. A Taxonomy of Multiple-Choice Item Writing. *Applied Measurement in Education*. *2*(1) 37-50.

Ningrum, M. N., Novi, R. D., & Parmin. 2018. Pengembangan Modul *Pop-up* Berbasis Inkuiri Terbimbing pada Tema Tata Surya untuk Kelas VII SMP. *Jurnal Inovasi Pendidikan IPA. 4*(1) 1-10.

Oller, J.W., & Perkins, K. 1978. *Language in Education: Testing the Tests*. Rowley, Mass.: Newbury House Publishers, Inc.

Permendikbud Nomor 104 Tahun 2014 tentang Penilaian Hasil Belajar Oleh Pendidik Pada Pendidikan Dasar dan Pendidikan Menengah.

Pramana, W. D. & Novi, R. D. 2014. Pengembangan *E-Book* IPA Terpadu Tema Suhu dan Pengukuran untuk Menumbuhkan Kemandirian Belajar Siswa. *Unnes Science Education Journal. 3*(3) 602-608.

Ramirez., Rachel, P. B., & Mildred, S.G. 2008. *Creative Activities and Student's Higher Order Thinking Skills*. Filipina: U. P. College of Education.

Rusilowati, A., Supriyadi., & Arif, W. 2015. Pembelajaran Kebencanaan Alam Bervisi SETS Terintegrasi dalam Mata Pelajaran Fisika Berbasis Kearifan Lokal. *Jurnal Pendidikan Fisika Indonesia*. *11*(1) 42-48.

Setyanto, H., Sudarmin., & Novi, R. D. 2015. Pengembangan LKS IPA Berbasis *Problem Based Learning* pada Tema Pencemaran Lingkungan Guna Menumbuhkan Kemandirian Siswa. *Unnes Science Education Journal. 4*(3) 990-997.

Slameto. 2010. *Belajar dan Faktor yang Mempengaruhinya*. Jakarta: Rineka Cipta.

Sugiyono. 2012. *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.

Treagust, D. F. 2006. *Diagnostic Assessment in Science as a Means to Improving Teaching, Learning and Retention*. UniServe Science Assessment Symposium Proceedings.

Tuysuz, C. 2009. Development of Two-Tier Diagnostic Instrument and Assess Student's Misunderstanding in Chemistry. *Scientific Research and Essay*. *4*(6) 626-631.

Yaumi, M. 2013. *Prinsip-Prinsip Desain Pembelajaran*. Jakarta: Kencana Prenada Media Group.

Zainuddin. 2016. Meningkatkan Kemampuan Berpikir Tigkat Tinggi Mahasiswa pada Perkuliahan Ekspeerimen Fisika I Melalui Penerapan Model Inquiry Discovery Learning. *Prosiding Seminar Nasional Pendidikan IPA.* Banjarmasin: S2 IPA UNLAM PRESS.