Ilkogretim Online - Elementary Education Online, Year; Vol 20 (Issue 5): pp. 122-128
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.12

ANALYZING THE INSTRUMENT USED TO MEASURE STUDENTS' HIGHER ORDER THINKING SKILL IN ENVIRONMENTAL EDUCATION LEARNING

R. Sihadi Darmo Wihardjo, Environmental Education, Graduate Program, Universitas Negeri Jakarta, Jakarta, Indonesia, rsihadi@unj.ac.id

Abstract - This research aims to analyze the feasibility and validity of the assessment instrument used to measure students' Higher Order Thinking Skill (HOTS) in Environmental education learning. The Borg and Gall development model was modified in the following stages: (1) Information collection, (2) Making plans, (3) Form preparation, (4) Revision, and (5) Product implementation. Data were obtained from 134 students of the population and Environmental Education Study Program at Jakarta State University, using 25 question items in the form of HOTS multiple choices. Meanwhile, QUEST was the analysis technique used to examine the validity, reliability, difficulty level, and questions' differentiator of the data. The result showed that the features of the generated assessment instrument are feasible and valid as alternative in assessing students' HOTS.

Keywords: HOTS, Environmental education learning, assessment instrument

I. INTRODUCTION

Nowadays, many students do not to possess higher-order thinking skills (HOTS),therefore, this study aims at developing assessment instruments to reform their knowledge. The Bloom concept in a book titled "Taxonomy of Educational Objectives: the classification of Educational Goals" was first used to determine HOTS(Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). This book was used to agglomerate the Bloom's Taxonomy, from the lowest to the highest thinking level.

According to researchAnderson et al. (1992), this concept is divided into three aspects, namely 1) Cognitive which is a mental action used to acquire knowledge, 2) Affective which is based on emotional attitudes and feelings, and 3) Psychomotor which is the physical ability used to perform a task. The learning and teaching process is indispensable following the current development and evolutionary demand of education (Chiu, 2016; Ramadhan, Mardapi, Prasetyo, & Utomo, 2019). HOTS is also defined as a tool used to facilitate the thinking process with many variables, such as in guiding and encouragingt students to achieve set goals (Ramadhan, Sumiharsono, Mardapi, & Prasetyo, 2020).

Istiyono, Mardapi, and Suparno (2014) reveals that increasing the ability of students' Higher Order Thinking skills (HOTS) can be done through the appropriate learning and assessment models. The statement indicates the importance of appropriate learning and assessment models in improving HOTS capabilities. Having the right assessment can encourage students to learn by thinking high(Afflerbach, Cho, & Kim, 2015; Ramadhan, Sunarto, Mardapi, & Prasetyo, 2020). Thus, an appropriate assessment is to not only measure Lower Order Thinking Skill (LOT), which includes students' ability to remember or understand, but also the Higher Order Thinking Skill (HOT) which includes students ' ability to analyse, evaluate and create.

High Order thinking, always associated with Bloom's taxonomy especially the top three levels of thinking in Bloom's taxonomy, namely synthesis (C4), analysis (C5) and evaluation (C6) or in accordance with the revisions of Anderson and Krathwool are analysis (C4), Evaluation (C5) and creating (C6)(Istiyono et al., 2014; Ramadhan, Sunarto, et al., 2020). In fact, there are still many teachers who are not yet familiar with the High Order Thinking-based questions. In fact, the assessment should be using the Higher Order Thinking has begun to be introduced in the assessment process in class by the teacher. So that the ability to design and develop problems of HOTs must be owned by the teacher(Ramadhan, Sumiharsono, et al., 2020). If the ability of the teacher to design the HOTs level is only mediocre, this will be based on the lack of quality problems produced, and this will adversely affect the measurement and assessment process of the learners. The statement above is in line with the statement from Ong, Hart, and Chen (2016) stating how important the role of the teacher is to help students build their scientific ideas and reflective thinking skills of the students.

The teacher is fully responsible for students Environmental education learning at schools using books and lectures(Kışoğlu, Gürbüz, Erkol, Akar, & Akıllı, 2017). However, most students tend to lack enthusiasm, spend hours without understanding, become sleepy and lose focus. Therefore, they need to be involved in many activities that are supposed to be carried out teachers to achieve a conducive learning atmosphere. One of the main problems associated with students education is the technique used in dealing with the learning process, which focuses on the teacher as a transformer, involve in many development processes. This tends to enable students to analyze, evaluate and create each lesson properly.

The above descriptions are the background of the study used to develop multiple-choice HOTS items. The material used comprises of "Global climate change Issues", and it was chosen due to its relevance to everyday life. This study therefore has the ability to help students analyze, evaluate and create a sophisticated and progressive thinking transformation beneficial to society in future.

II. LITERATURE REVIEW

Assessment Instrument

The instrument is defined as a tool used to determine the academic requirements of an object, its accuracy, validity, and reliability on students, curriculum, programs and educational policies (Brookhart & Nitko, 2008; Satria & Uno, 2012). According to (Mardapi, 2008), there are two types of assessment instruments, namely test and non-test. In the educational framework, the tests used to measure the achievement, intelligence, talents, and skills of students, while non-tests are for attitudes, observations, and guidelines(Gardner, 2006; Gikandi, Morrow, & Davis, 2011; Mardapi, 2004, 2008). The assessment functions as follows (1) a tool used to determine the instructional objectives achieved, (2) a feedback tool used to improve the teaching and learning process. The improvements might be implemented in instructional activities, teaching strategies, etc and (3) a report used in learning many science subjects(Mardapi, 2012).

Higher-Order Thinking Skill (HOTS)

According to King, Goodson, and Rohani (1998) this is a selective, creative, logical, critical, and meta-cognitive thinking process used to implement the thinking concept when students have difficulties. Brookhart (2010)defined it as the analysis, evaluation, creation, logic, and ability to think critically while solving problems.

Bloom developed a thinking concept called Bloom's Taxonomy, which consists of synthesis analysis (C4), evaluation (C5) and creativity (C6). It also consists of low order thinking skills or LOTS which involve the recitation ability (C1), understand (C2), and implement (C3) (Anderson et al., 1992; Anderson et al., 2001).

In line with Bloom's taxonomy, Ramadhan, Mardapi, Prasetyo, et al. (2019) stated that the recitation ability (C1) is limited to repeating past events, understanding (C2) comprises of absorption of information, while interpretation is associated with exploration. Implementing (C3), is used to generalize a situation that has been previously described, analyzing (C4) connects with one another systematically and in a structured manner, with the ability to solve problems through facts. Evaluation (C5) means conducting an assessment based on criteria or standards, while creativity (C6) is the highest level of HOTS where students have the problem-solving ability through creative thinking level.

Based on these experts, it can be concluded that high-level thinking is the ability to use complex, critical, creative and solutiphic thinking in resolving problems that have never been found before or different from examples. An issue or problem that was originally in the HOTs category could be no longer HOTs if it had been delivered before and the student finally recorded the problem as a memory or memorization. All these skills are active when a person is dealing with unusual problems, uncertainties, questions and choices

Multiple choice

Multiple-choice tests are objective in large and small-scale tests such as Formative, and Summative Test. According to Gronlund and Linn (1990), multiple-choice questions are used as the subject matter in measuring complex thinking. They are also used as parameters to determine the causes of high difficulty level, such as the existence of a distractor.

HOTS issues generally prioritize the insertion of stimulus in contextual situations. The answer key is not explicitly contained in the reading or stimulus, as respondents utilize the questions and background knowledge background, to state the reasons. The complexity of multiple-choice questions is to comprehensively test students' understanding of a problem related to one statement. Similar to multiple choice questions, HOTS comprises of stimulus based on the contextual situations.

III. RESEARCH METHODOLOGY

This study utilized the Research and Development method to develop HOTS items for students of population and Environmental Education Study Program at Jakarta State University. The procedures in this researchwere adapted from development steps by Gall, Borg, and Gall (1996), which were modified into the following stages: (1) Information collection, (2) Making plans, (3) Form preparation, (4) Conducting revision, and (5) Product implementation.

The HOTS assessment instrument in this study used 25 multiple choice questions based on the topic "Global Climate change Issues." The test instrument was distributed to134respondents using previously studied material on contextual cases. The descriptive analysis techniques were used to process the data conducted from the results of limited trials in the field. Meanwhile, the QUEST program was used to measure the validity, reliability, level of difficulty and differentiation of items. The validity results were conducted through MNSQ INFIT analysis and Item fit, with the Rasch model used to facilitate the interpretation of statistical reliability test results. The distinguishment power analysis uses the biserial point value in the Quest program.

The item match test on this study uses the model Rasch assumption. Grain match Test Assuming model Rasch approach is done by viewing the fit or absence of grain against the model. This test is analyzed by using the Winstep Program. Terms of grain is said to be fit against the model of Winstep program among others if the value of Outfit MNSQ of 0.5 to 1.5 and Outfit ZSTD value of-2 to 2, as well as Pt-measure Corr positive value then it can be said the item is fit or suitable against the model. The item is said to be fit when fulfilled one of the three conditions.

Analysis of difficulty level of test device is done by using computer program MicroCat Iteman. The problem difficulty level can be seen in the Prop column. Correct. The problem that has a good difficulty level is in the interval 0.3 to 0.8. Analysis of the different power of test devices can be seen in the Biserial Point field conducted using MicroCat Iteman computer program. Criterion criteria good thing has a value of $D \geq 0.3$, while the problem that has a value $D \leq 0.3$ should be revised or replaced with a new problem(Mardapi, 2012; Ramadhan, Mardapi, Sahabuddin, & Sumiharsono, 2019).

IV. RESULTS AND DISCUSSION

A total of two initial stages were carried out before starting the test. In the first stage, the instrument was assessed by several instruments and product expert, as well as 2 material experts. Meanwhile, in the second stage, 134 students were used in a trial test on multiple-choice HOTS questions that had passed the expert validation test.

The Quest Program is an item analysis application developed based on applied statistics and based on the item response theory, also known as Latent Trait Theori (LTT) or Characteristics Curve Theory (CCT). There are two postulates as the basis of item response theory. The first is a set of factors, namely latent, verbal, psychomotor, and cognitive traits. The second postulate is the item characteristics curve (ICC) consisting of respondents and item sets. The logistics which is studied in PMM activities are a one-parameter logistic model (rasch model), or 1-parameter logistic response theory (IRT 1-PL) used to analyze data that focuses on the level of difficult parameters.

Adams and Khoo (1996)stated that Quest analyzes items, defines a participant's ability = Θ and the difficulty level of an item. The Rasch Model is a central element, or one parameter (1-PL). Itanal in the syntax section is output command on the statistics of test on difficulty, discrimination, and distractor levels. The output provides information on item statistics and test kits such as the degree of difficulty and discriminatory power. Meanwhile, the Quest analyzes respondents that are judged dichotomically (1-10) or politically (1-2-3-4-etc.). Unconditional (UCON) or joint maximum like the hood is used by Quest to estimate the subject and measure the validity, reliability, level of difficulty and differentiation.

Results of Limited Test Data Validity

The good learning outcomes are valid results tests (Ramadhan, Mardapi, Prasetyo, et al., 2019; Ramadhan, Sumiharsono, et al., 2020). In this study, the researcher used limited trials carried out on134 students. The multiple-choice HOTS question was used for 60 minutes and a trial.

In addition, the validity results were obtained through MNSQ INFIT analysis. The problem is declared valid, assuming it is in the range of -2.0 to +2.0 with the FIT statement. However, after analysis results, 25 items were declared fit. Table 1 shows the results of the validity of questions using INFIT analysis on MNSQ data from 134 students of population and Environmental Education Study Program at Jakarta State University.

Table 1: Problem multiple-choice HOTS declared Valid

Item No.	INFIT MNSQ	Description
1	0,91	FIT
2	0,99	FIT
3	1,03	FIT
4	1,07	FIT
5	0,99	FIT
6	0,88	FIT
7	1,12	FIT
8	0,98	FIT
9	1,02	FIT
10	0,98	FIT
11	0,84	FIT
12	0,96	FIT
13	0,96	FIT
14	1,15	FIT
15	0,92	FIT
16	1,05	FIT
17	0,87	FIT
18	0,99	FIT
19	1,15	FIT
20	0,92	FIT
21	1,00	FIT
22	0,85	FIT
23	0,98	FIT
24	1,23	FIT
25	0,98	FIT

Item Reliability Analysis Problem

Reliability is a measuring tool used to determine the quality of an item. A test is termed reliable when the same result is obtained at different times in groups. In addition, a measurement is called stabilitywhen the conditions and opportunities have the same result (Mardapi, 2008).

The analysis of Item fit is valid when in the range of 0.77 to 1.30. The reliability value of 0.87 shows that the questions are in the high category which means that the test instrument is reliable, However, due to its high level of the reliability coefficient of education, the questions are not very good. The average level of compatibility of the items is 1.0, and the standard deviation is 1.11, therefore, overall, the respondents are suitable with the model set of Rasch.

Item Difficulty Level Analysis

Boopathiraj and Chellamani (2013)stated thatthe item difficulty is the proportion of respondents that correctly mark items. Good questions are items with mediocre difficulty and with answering difficulty.

Table 2 Results of difficulty level output of the Quest program

Item No.	*Threshold*	Categories
1	0,78	Moderate
2	0,55	Moderate
3	0,41	Moderate
4	0,41	Moderate

5	0,18	Difficult
6	0,61	Moderate
7	0,73	Moderate
8	-0,19	Difficult
9	0,81	Easy
10	0,59	Moderate
11	0,69	Moderate
12	0,41	Moderate
13	0,41	Moderate
14	1,10	Easy
15	1,59	Easy
16	0,55	Moderate
17	0,58	Moderate
18	0,61	Moderate
19	0,55	Moderate
20	0,71	Moderate
21	0,48	Moderate
22	0,04	Difficult
23	0,14	Difficult
24	0,66	Moderate
25	0,10	Difficult

Table 2 showed a total of 5, 17, and 3 problem in difficult, moderate and easyCategories, of20%, 68% and 12%.

The Analysis of Distinguished Items

According to (Mardapi, 2008), irrespective of the ability of an item to distinguished students with low or high capabilityin problem analysis,there are characteristics of the positive sign of discrimination index. Students in this category are in the smart category and answer more questions correctly. The item is said to have no distinguishing ability with symbol D equals 0. This means that students in both Upper and Lower groups answered the questions correctly.

Table 3: Results of distinguishing power using Biseral points

Item No.	*Point Biserial* (ρ_{bis})	Categories
1	0,40	Good
2	0,18	Not good
3	0,53	Good
4	0,24	Enough
5	0,20	Enough
6	0,49	Good
7	0,40	Good
8	0,22	Enough
9	0,08	Not good
10	0,40	Good
11	0,40	Good
12	0,52	Good
13	0,30	Good
14	0,33	Good
15	0,35	Good
16	0,40	Good
17	0,35	Good
18	0,40	Good
19	0,49	Good
20	0,40	Good
21	0,52	Good
22	0,28	Enough
23	0,22	Enough
24	-0,06	Not good
25	0,40	Good

Table 3 shows 17, 5 and 3 quality questions in good, Enough and not good categories, respectively. It means that majority of questions are acceptable and can be implemented on the students.

Product Revision

The valid and reliable criteria are carried out to obtain the final product revision. The validation and product revisions on limited trial are based on assessments. The average HOTS test questions on the Basic consists of 25 feasible and valid questions. Generally, the insights and suggestions from the validator possess a better version of the language, produce questions, material focus, and material sequence.

The research was conducted to develop diagnostic tests that can be used to measure the high level of thinking skills of students. In Indonesia, high level thinking is always associated with Bloom's taxonomy revision, especially the top three, namely C4 (analyze), C5 (Evaluate) and C6 (create). Even the use of bloom taxonomy is also contained in the curriculum used in Indonesia. Therefore, the concept of high-level thinking used in this research refers to the concept of high-level thinking of Bloom's revision.

Proof by using classical test theory with consideration of difficulty level, and different power. Grain difficulty level in order to be acceptable when the magnitude of 0.30 to 0.80, the power of grain in order to be acceptable is when the magnitude of 0.30. Based on empirical test results with classical test theory, there is information that there are 3 problems that are accepted with the terms and 23 items received. The reason for acceptance by condition is to be used for a different power value less than 0.3. The problem is received with these conditions later in the revision again to produce quality problems, but overall there is no problem of bad or waste. Results of analysis by using Iteman obtained reliability value of 0.87. The value is quite large and indicates that the instrument is reliable. Empirical validity using the Rasch model has also been qualified. As such, this instrument is qualified to be used to measure the ability of high-level thinking, based on the topic "Global Climate change Issues".

Final Product Review

The HOTS assessment instrument for studentson the competence of “Global Climate change Issues" is the final result of this study. In addition, the developed multiple-choice questions have passed limited trials, with the instrument, product and material experts involved in perfecting this product. The improvements were made after getting the results from validation and limited trials. The product developed has met the criteria of a decent item with the quality of items has been tested through validation, reliability, level of difficulty, and distinguishing features.

V. CONCLUSION

Thefeasibility and validityof the assessment instrument used to measure students’ Higher Order Thinking Skill (HOTS) in Environmental education learningled to the following conclusion: (1)The multiple choice of HOTS as an instrument provided five options; in accordance with Global Climate change Issues” for Graduate student of population and Environmental Education Study Program at Jakarta State University. (2) The validity of HOTS questions based on validator analysis consists of the instrument, product, and material experts, which showed that HOTS assessment instrument is feasible in terms of validity, reliability, difficulty, and question differentiator used at schools and (3) Characteristics of 25multiple choice HOTS questions show that the quality of the question obtained from the item analysis was valid.

REFERENCES

1. Adams, R. J., & Khoo, S.-t. (1996). Quest: The interactive test analysis system, Version 2.1. Melbourne: ACER.
2. Afflerbach, P., Cho, B.-Y., & Kim, J.-Y. (2015). Conceptualizing and assessing higher-order thinking in reading. Theory into practice, 54(3), 203-212.
3. Anderson, L. W., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., & Pintrich, P. (1992). A taxonomy for learning, teaching and assessing: A revision of Bloom’s taxonomy. New York: Longman Publishing. .
4. Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., . . . Wittrock, M. (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom’s taxonomy. New York: Longman Publishing.

5. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives, handbook I: The cognitive domain. New York: David McKay Co Inc.
6. Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. International journal of social science & interdisciplinary research, 2(2), 189-193.
7. Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom: ASCD.
8. Brookhart, S. M., & Nitko, A. J. (2008). Assessment and grading in classrooms: Prentice Hall.
9. Chiu, M.-S. (2016). The Challenge of Learning Physics Before Mathematics: A Case Study of Curriculum Change in Taiwan. Research in Science Education, 46(6), 767-786.
10. Gall, M. D., Borg, R., & Gall, P. (1996). Educational research: An instruction. New York: White Plains: Longman.
11. Gardner, J. (2006). Assessment and learning: Sage.
12. Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. Computers & education, 57(4), 2333-2351.
13. Gronlund, N., & Linn, R. (1990). Measurement and Evaluation in Teaching 6th Ed. USA: Millan Publishing Company.
14. Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA. Jurnal Penelitian dan Evaluasi Pendidikan, 18(1), 1-12.
15. King, F., Goodson, L., & Rohani, F. (1998). Higher order thinking skills. Retrieved January, 31, 2011.
16. Kışoğlu, M., Gürbüz, H., Erkol, M., Akar, M. S., & Akıllı, M. (2017). Prospective Turkish elementary science teachers' knowledge level about the greenhouse effect and their views on environmental education in university. International Electronic Journal of Elementary Education, 2(2), 217-236.
17. Mardapi, D. (2004). Penyusunan tes hasil belajar. Yogyakarta: PPS UNY.
18. Mardapi, D. (2008). Teknik penyusunan instrumen tes dan nontes. Jogjakarta: Mitra Cendekia.
19. Mardapi, D. (2012). Pengukuran penilaian dan evaluasi pendidikan. Yogyakarta: Nuha Medika.
20. Ong, K. K. A., Hart, C. E., & Chen, P. K. (2016). Promoting Higher-Order Thinking Through Teacher Questioning: a Case Study of a Singapore Science Classroom. New Waves, 19(1), 1.
21. Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The Development of an Instrument to Measure the Higher Order Thinking Skill in Physics. European Journal of Educational Research, 8(3), 743-751. doi: 10.12973/eu-jer.8.3.743
22. Ramadhan, S., Mardapi, D., Sahabuddin, C., & Sumiharsono, R. (2019). The Estimation of Standard Error Measurement of Physics Final Examination at Senior High Schools in Bima Regency Indonesia. Universal Journal of Educational Research, 7(7), 1590 - 1594. doi: 10.13189/ujer.2019.070713
23. Ramadhan, S., Sumiharsono, R., Mardapi, D., & Prasetyo, Z. K. (2020). The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis. International Journal of Instruction, 13(2). doi: 10.29333/IJI.2020.13235A
24. Ramadhan, S., Sunarto, Mardapi, D., & Prasetyo, Z. K. (2020). Higher Order Thinking Skill in Physics; A Sistimatical Review. International Journal of Advanced Science and Technology, 29(05), 5102 - 5112.
25. Satria, K., & Uno, H. B. (2012). Assesment Pembelajaran. Jakarta: Bumi Aksara.