The Analysis of Instrument Quality to Measure Graduate Students' Higher Order Thinking Skill in Environmental Education Learning

R. Sihadi Darmo Wihardjo[1]*, Ayub Muktiono[2], Syahrul Ramadhan[3], Ibnu Salman[4], Deni Hadiana[5].

[1,2]*Universitas Krisnadwipayana, Indonesia,* E-mail: *sihadiwihardjo@gmail.com*

[3,4,5]*National Research and Innovation Agency, Jakarta, Indonesia*

Abstract: This research aims to develop a valid and reliable assessment instrument for measuring students' Higher Order Thinking Skills (HOTS) in environmental education learning. The study follows a research and development design, adapted from the development model proposed by Borg and Gall. The researcher modified the original model into several stages, including (1) gathering information, (2) making a plan, (3) preparing the product form, (4) conducting revisions of the initial product, and (5) implementing the final product. The trial subjects for this research consisted of 34 graduate students enrolled in the Population and Environmental Education Study Program at Jakarta State University. Data collection involved administering a test comprising 20 multiple-choice items related to environmental education learning, specifically focusing on HOTS. The QUEST program was utilized to analyze the collected data, examining the validity, reliability, difficulty level, and question differentiator of the assessment instrument. The results of this research demonstrate the successful development of a feasible assessment instrument for measuring graduate students' HOTS. The instrument exhibits high validity and reliability, ensuring that it accurately measures the targeted higher-order thinking skills. Moreover, the difficulty level and question differentiator of the items are well-balanced, allowing for effective differentiation among students' thinking abilities. This assessment instrument holds significant potential as an alternative tool for evaluating graduate students' HOTS in environmental education learning. It provides educators with a reliable means to assess and monitor students' higher-order thinking abilities, enabling the identification of areas that require further development. By incorporating this instrument into the assessment process, educators can facilitate the cultivation of critical thinking, problem-solving, and creativity skills among graduate students.

Keywords: HOTS, Environmental Education Learning, Assessment Instrument

1. INTRODUCTION

In contemporary times, the inadequate development of higher order thinking skills (HOTS) among students has become a pressing concern, necessitating the reform of existing learning systems. The origins of HOTS can be traced back to the seminal work by Bloom, as outlined in his renowned book titled "Taxonomy of Educational Objectives: The Classification of Educational Goals" [1]. This taxonomy encompasses a comprehensive framework that categorizes various levels of thinking, progressing from the lowest to the highest order.

Bloom's concept of learning goals is divided into three primary domains: Cognitive (pertaining to mental actions in acquiring knowledge), Affective (relating to attitudes and emotions), and Psychomotor (concerning physical abilities and skills) [2]. In line with the progressive and evolving demands of education, the cultivation of higher order thinking skills (HOTS) is considered indispensable in the realm of teaching and learning [3]. HOTS serves as a powerful tool that facilitates the thinking process, influenced by numerous variables contingent upon specific contextual conditions. It is widely acknowledged that providing guidance and encouragement to students is crucial for their attainment of HOTS objectives [4].

Within the realm of environmental education, teachers and educators play a pivotal role. As the primary sources of information, educators often rely on traditional teaching methods, such as recitation, centered around textbooks and lectures. Unfortunately, this approach frequently leads to student disengagement, apathy, drowsiness, and diminished focus over prolonged periods. To create an optimal learning environment, educators should employ a variety of activities that foster student involvement, creativity, and active participation. Addressing this challenge necessitates a shift in the learning process, away from a teacher-centric model and toward an approach that actively involves

students in developmental processes, enabling them to become proficient in analyzing, evaluating, and creating knowledge in each lesson.

The aforementioned factors serve as the underlying motivation for the present study, which aims to develop an assessment instrument for Higher Order Thinking Skills, focusing specifically on the feasibility and validity of multiple-choice items. The chosen subject matter for this instrument development is "Global Climate Change Issues." This topic is selected due to its relevance to everyday life and its prominence in discussions and discourse surrounding environmental education. Furthermore, it is anticipated that this study will equip students with the ability to engage in sophisticated and progressive thinking, empowering them to contribute meaningfully to society in the future.

2. LITERATURE REVIEW

2.1. Assessment Instrument

An assessment instrument is a tool that fulfills academic requirements and allows for the measurement of specific objects, enabling accurate and reliable data collection. Assessment, on the other hand, refers to the process of gathering information to make informed decisions about students, curriculum, programs, and educational policies [5]. According to Mardapi [6], assessment instruments can be categorized into two types: tests and non-tests. Tests are utilized to measure learning achievements, intelligence, talents, and skills of students, while non-tests encompass assessments of attitudes, observations, and guidelines [6, 7]. The primary functions of assessment include: (1) determining whether instructional objectives have been achieved, which necessitates alignment with instructional formulations; (2) providing feedback to enhance the teaching and learning process, with improvements implemented in instructional activities, student learning activities, and teacher teaching strategies; and (3) reporting on learning progress in various subjects to students' parents.

2.2. Higher Order Thinking Skills (HOTS)

Higher Order Thinking Skills, as elucidated by King et al. [8], encompass the selective, creative, logical, critical, and metacognitive processes of thinking. These skills are particularly relevant when students encounter challenges. Brookhart [9] defines HOTS in terms of analysis, evaluation, creation, logical reasoning, critical thinking, problem-solving abilities, and creative thinking.

These thinking concepts are developed within Bloom's Taxonomy. According to Bloom, cognitive processes consist of HOTS skills, which involve synthesis analysis (C4), evaluation (C5), and creation or creativity (C6), as well as lower-order thinking skills (LOTS) that encompass recitation (C1), understanding (C2), and implementation (C3) (Anderson and Krathworl, 2001: 68-88). Kusnawa [12] further expands on Bloom's taxonomy, explaining that recitation (C1) is limited to repeating past events, understanding (C2) entails absorbing information, interpreting meaning, and exploring, implementing (C3) involves generalizing a previously described situation, analyzing (C4) involves systematically connecting information and problem-solving based on facts, evaluating (C5) entails conducting assessments based on criteria or standards, and creating (C6) represents the highest level of HOTS, where students demonstrate problem-solving abilities through creative thinking.

2.3. Multiple Choice

Multiple-choice tests are widely used in both large-scale and small-scale assessments, such as formative and summative tests. They offer the advantage of relatively easy scoring. Gronlund and Linn [10] suggest that multiple-choice questions can be adjusted to measure thinking skills ranging from simple to complex, depending on the subject matter. Multiple-choice questions often include distractors, which contribute to their high difficulty level.

HOTS-related issues typically emphasize the incorporation of stimuli within contextual situations. The answer key is not explicitly provided in the reading or stimulus, requiring respondents to draw on their background knowledge and provide reasoning for their answers. The complexity of multiple-choice

questions lies in the comprehensive testing of students' understanding of a problem and its interconnectedness to other statements. Similarly, HOTS questions in the form of multiple choices include stimuli based on contextual situations.

3. METHOD

3.1. Research Design

This study follows a Research and Development (R&D) approach with the objective of developing Higher Order Thinking Skills (HOTS) items for the graduate student population in the Environmental Education Study Program at Jakarta State University. The research development process incorporates the steps outlined by Gall et al. [11]. However, the researcher has modified the steps into several stages, which include: (1) Gathering Information; (2) Making a Plan; (3) Preparing Product Forms; (4) Conducting Initial Product Revisions; and (5) Product Implementation.

3.2. Sample

The sample for this study consists of 34 respondents who are enrolled in the graduate program of the Environmental Education Study Program at Jakarta State University. These participants were chosen based on their availability and willingness to participate in the study.

3.3. Instrument

The HOTS assessment instrument used in this study consists of 20 multiple-choice questions. The questions are centered around the topic of "Global Climate Change Issues." The selection of this topic is based on its relevance to environmental education and its significance in contemporary discussions. The test instrument was administered to the participants, incorporating both materials they had previously studied and contextual cases related to the topic.

3.4. Data Analysis Technique

Descriptive analysis techniques were employed to process the data collected from the limited trials conducted in the field. To measure the validity, reliability, level of difficulty, and item differentiation, the Quest program was utilized. The validity of the instrument was assessed through MNSQ INFIT analysis and Item fit, providing insights into the conformity of the instrument items with the intended construct. The reliability test of the instrument in this study utilized the Rasch model, which offers a statistical interpretation of the instrument's reliability. The item differentiation analysis employed the biserial point value within the Quest program, providing information on the ability of the items to discriminate between respondents.

By employing these comprehensive data analysis techniques, this study aims to ensure the quality and accuracy of the developed HOTS assessment instrument for measuring the higher order thinking skills of graduate students in the field of environmental education.

4. RESULTS

There are two stages that must be done before starting the test. At the first stage, the instrument was assessed by several experts, consisting of 1 instrument expert, 1 product expert , and 2 material experts. The second stage, 34 Graduate student of population and Environmental Education Study Program at Jakarta State University 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 based on a theory namely item response. Modern measurement theory is used in item analysis. Latent Trait Theori (LTT) or Characteristics Curve Theory (CCT) is another name for item response theory. There are two postulates as the basic of item response theory. The first one is a set of factors namely traits,

latent traits or abilities that can predict the ability of the subject. Verbal abilities, psychomotor abilities, cognitive abilities are called by traits . The second postulate is the item characteristics curve (ICC) which has the latent ability of respondents and item sets. The logistics model is studied in PMM activities, named : a one-parameter logistic model (rasch model) or 1-parameter logistic response theory (IRT 1-PL) to analyze data that focuses on the level of difficulty parameters.

Adams and Khoo [12] stated that Quest can analyze items. The Rasch Model is a central element, one parameter (1-PL). The Quest Program is a participant's ability = Θ□and the difficulty level of item b as the main item. Itanal in the syntax section is output command on the statistics of test on difficulty level, discrimination level, and distractor level. The output provides information about item statistics and test kits such as the degree of difficulty and discriminatory power. Quest analyzes respondents who are judged dichotomically (1-10) or politically (1-2-3-4-etc.). Unconditional (UCON) or joint maximum likehood is used by Quest to estimate the subject. The Quest program is used to be able to measure the validity, reliability, level of difficulty and differentiation of questions.

4.1. Results of Limited Test Data Validity

The good learning outcomes are valid results tests [3,4] . Limited trials were conducted from 34 Graduate student of population and Environmental Education Study Program at Jakarta State University. The multiple choice HOTS question in a limited trial is conducted in 60 minutes and one trial only.

The validity results were obtained through MNSQ INFIT analysis and Item fit. The problem is declared valid if it is in the range of -2.0 to +2.0 with the FIT statement. After analysis results, 20 items were declared fit. Here are the results of the validity of the questions using INFIT analysis of MNSQ data from 34 Graduate student of population and Environmental Education Study Program at Jakarta State University.

Table 1 Problem multiple choice HOTS declared Valid

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

4.2. Item Reliability Analysis Problem

Reliability is a measuring tool to determine a quality of item. A test is reliable if it is tested on the same group at different times. The measurement of stability is tested in many conditions and opportunities must have the same result [6]. The analysis of Item fit if it is in the range of 0.77 to 1.30 then items are valid. The questions made by writer were valid. The reliability value shows that the questions reliability in high category that is 0.87. It means that the test instrument is reliable, but it is still not very good due to its high level of the reliability coefficient of education. The average level of compatibility of the items is 1.0 and the standard deviation is 1.11, so overall the respondents are suitable with the model set of Rasch .

4.3. Item Difficulty Level Analysis

Boopathiraj and Chellamani [13] define that item difficulty is the proportion of respondents who correctly mark items. Items with mediocre difficulty and not easily answer are good questions.

Table 2 Results of difficulty level output of the Quest program

Item	*Threshold*	Kriteria
1	-0,88	Difficult
2	-1,21	Difficult
3	1,41	Easy
4	1,59	Easy
5	0,55	Moderate
6	-0,88	Difficult
7	-1,21	Difficult
8	0,81	Easy
9	-1,40	Difficult
10	1,41	Easy
11	1,41	Easy
12	0,18	Difficult
13	1,10	Easy
14	-0,73	Difficult
15	-0,19	Difficult
16	0,81	Easy
17	0,81	Easy
18	-1,04	Difficult
19	0,06	Difficult
20	-2,61	Difficult

Based on the data in table 1 above, 11 problems in difficult level with total score of 55%, questions with an easy level of 40% were 8 items, and 5% for questions with modiocre level was 1.

4.4. The Analysis of Distinguished Items

According to Mardapi [6], whether or not an item is able to distinguished students who have low or high ability is one of problem analysis objectives. The characteristics of categorize ability are having positive sign of discrimination index. Students in this category are smart students. Students in the smart category answer more questions correctly. The item is said to have no distinguishing ability at all with symbol D = 0. It means that both of the Upper group students and Lower group students answered correctly.

Table 3 Results of distinguishing power using Biseral points

Item	*Point Biserial* (ρ_{bis})	Kriteria
1	0,08	Not good
2	0,40	Good
3	0,28	Enough
4	0,30	Good
5	0,33	Good
6	0,35	Good
7	0,18	Not good
8	0,18	Not good
9	0,24	Enough
10	0,53	Good
11	0,20	Enough
12	0,49	Good
13	0,08	Not good
14	0,26	Not good
15	0,10	Not good
16	0,49	Good
17	0,52	Good
18	0,28	Enough
19	-0,06	Not good
20	0,22	Enough

The data in table 2 shows 8 good quality questions, 7 poor quality questions and 5 mediocre quality questions. It means that the questions made by writer is accepted because the majority of questions are acceptable and can be implemented on students.

4.5. Product Revision

The valid and reliable criteria are conducted to gain the final product in product revision. The validation revision and product revision on limited trial are the product revision of this study that based on product trial assessment. The average HOTS test questions on the Basic Competence "Global Climate change Issues" which consists of 20 HOTS questions are feasible and valid. Generally, the insights and suggestions from validator to be a better version are the language, produce questions, material focus, and material sequence.

4.6. Final Product Review

The HOTS assessment instrument for Graduate student of population and Environmental Education Study Program at Jakarta State University, Competence “Global Climate change Issues" is the final result of this study. HOTS multiple choice questions developed have passed trials in limited trials. Instrument experts, product experts and material experts are involved in the process of perfecting this product. The improvements were made after getting the results from validator validation and limited trials. The product developed has met the criteria of a decent item. The quality of items has been tested through validation, reliability, level of difficulty and distinguishing features.

CONCLUSION

In conclusion, this study aimed to analyze the quality of the assessment instrument used to measure the Higher Order Thinking Skills (HOTS) of graduate students in the Population and Environmental Education Study Program at Jakarta State University. The instrument employed in this study consisted of multiple-choice questions that focused on the Basic Competence of "Global Climate Change Issues." The findings of the research and subsequent discussion lead to the following conclusions; 1) The assessment instrument developed for measuring HOTS in the context of environmental education learning proved to be effective for graduate students in the Population and Environmental Education Study Program. The instrument consisted of multiple-choice questions with five response options, allowing for a comprehensive evaluation of students' higher-order thinking skills. 2) The validity of the HOTS questions was confirmed through rigorous analysis conducted by instrument experts, product experts, and material experts. The results of the instrument expert analysis demonstrated that the HOTS assessment instrument possessed the necessary qualities of validity, reliability, appropriate difficulty levels, and effective question differentiation. These attributes make the instrument a viable alternative for assessing students' higher-order thinking skills in educational settings. 3) The characteristics of the multiple-choice HOTS questions revealed the quality of the question items as determined through item analysis. The calculation of HOTS question validity revealed that all 20 questions demonstrated validity, further affirming the instrument's ability to accurately measure the targeted higher-order thinking skills.

It is important to acknowledge the limitations of this study. Firstly, the research was conducted solely within the Population and Environmental Education Study Program at Jakarta State University, which may limit the generalizability of the findings to other contexts. Additionally, the sample size of the study was relatively small, consisting of 34 graduate students. A larger and more diverse sample would enhance the robustness of the results. Future research should consider expanding the study to include a broader range of participants from different educational institutions and backgrounds. This study opens up possibilities for further research in several areas. Firstly, future studies could explore the application of the HOTS assessment instrument in different academic disciplines and educational levels to assess the transferability of the instrument across various contexts. Moreover, investigating the effectiveness of instructional interventions designed to enhance higher-order thinking skills would provide valuable insights into instructional practices that foster critical thinking, problem-solving, and creativity. Finally, examining the relationship between students' higher-order thinking skills and their

academic achievement or real-world problem-solving abilities would contribute to a deeper understanding of the impact of HOTS development on overall learning outcomes.

Finally, the analysis of the instrument quality to measure graduate students' higher-order thinking skills in environmental education learning yields promising results, indicating the viability and effectiveness of the HOTS assessment instrument. Despite the limitations of the current study, the findings provide a foundation for future research endeavors aimed at refining the instrument and exploring its broader applicability in diverse educational contexts.

REFERENCES

[1] B. S. Bloom, "Taxonomy of educational objectives: The classification of educational goals," *Cognitive domain,* 1956.

[2] L. W. Anderson, D. R. Krathwohl, P. Airasian, K. Cruikshank, R. Mayer, P. Pintrich, *et al.*, "A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy," *New York. Longman Publishing. Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction,* vol. 9, pp. 137-175, 2001.

[3] S. Ramadhan, D. Mardapi, Z. K. Prasetyo, and H. B. Utomo, "The Development of an Instrument to Measure the Higher Order Thinking Skill in Physics," *European Journal of Educational Research,* vol. 8, pp. 743-751, 2019.

[4] S. Ramadhan, R. Sumiharsono, D. Mardapi, and Z. K. Prasetyo, "The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis," *International Journal of Instruction,* vol. 13, 2020.

[5] K. Satria and H. B. Uno, "Assesment Pembelajaran," *Jakarta: Bumi Aksara,* 2012.

[6] D. Mardapi, "Teknik penyusunan instrumen tes dan nontes," *Jogjakarta: Mitra Cendekia,* 2008.

[7] D. Mardapi, "Penyusunan tes hasil belajar," *Yogyakarta: PPS UNY,* 2004.

[8] F. King, L. Goodson, and F. Rohani, "Higher order thinking skills," *Retrieved January,* vol. 31, p. 2011, 1998.

[9] S. M. Brookhart, *How to assess higher-order thinking skills in your classroom*: ASCD, 2010.

[10] N. Gronlund and R. Linn, "Measurement and Evaluation in Teaching 6th Ed. USA: Mc," ed: Millan Publishing Company, 1990.

[11] M. D. Gall, R. Borg, and P. Gall, "Educational research: An instruction," *New York, White Plains: Longman,* 1996.

[12] R. J. Adams and S.-t. Khoo, "Quest: The interactive test analysis system, Version 2.1," *Computer software]. Melbourne: ACER,* 1996.

[13] C. Boopathiraj and K. Chellamani, "Analysis of test items on difficulty level and discrimination index in the test for research in education," *International journal of social science & interdisciplinary research,* vol. 2, pp. 189-193, 2013.

DOI: https://doi.org/10.15379/ijmst.v10i3.1528