Journal of Turkish Science Education, 2021, 18(4), 674-690.
DOI no: 10.36681/tused.2021.97

HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers

Ika Maryani[1], Zuhdan Kun Prasetyo[2], Insih Wilujeng[3], Siwi Purwanti[4], Meita Fitrianawati[5]

[1] *Universitas Negeri Yogyakarta,Yogyakarta, Indonesia, ORCID ID: 0000-0002-7154-2902*
[2] *Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia, ORCID ID: 0000-0001-9342-1565*
[3] *Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia, ORCID ID: 0000-0003-1900-7985*
[4] *Department of Elementary School Teacher Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, ORCID ID: 0000-0002-1433-7531*
[5] *Department of Elementary School Teacher Education, Universitas Ahmad Dahlan, Yogyakarta, Indonesia, ORCID ID: 0000-0002-3748-3718*

ABSTRACT
Higher-order thinking skills (HOTs) are very crucial thinking skills needed by teachers to train students to develop 21st-century learning. This study aimed to develop Multiple Choice and Essay Questions to measure the HOTs of the prospective teachers of the elementary school education department. This study used a 4-D model by Thiagarajan which involved experts at natural science, evaluation studies, and primary school pedagogy in the content validation. We also involved 156 prospective teachers as the test subjects. The assessment of instrument quality by experts showed that the question quality was very good. This research succeeded in developing 10 multiple choice questions and 5 essays. The validity test by Rasch Model showed that there were 7 multiple choice questions classified as fit, and 3 questions were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with KR-20 on multiple-choice questions and Cronbach's alpha for the essay questions resulted reliable questions. The discrimination index showed discarded, sufficient, good, and very good. The item difficulty index showed that 3 questions are moderate (num 7, 1, 5) and 7 questions are difficult (num 4, 10, 6, 3, 2, 8, 9). The distractor efficiency showed that 59.2% of distractors worked, and 40.8% did not work. This instrument can be used to analyze prospective teachers' HOTs. This data can be used as the reference for developing competency improvement programs for prospective teachers, for example through the HOTs-oriented learning models.

ARTICLE INFORMATION
Received: 26.07.2020
Accepted: 16.07.2021

KEYWORDS: Higher-order thinking skills, multiple choice, essay questions, instrument.

Introduction

21st-century education requires students to have life skills, such as innovative, creative, adaptive, and technology literate. Based on this change, an institute of teacher training is required to produce qualified prospective teachers. Bhakti and Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professionals who provide expert service and demand academic, pedagogical, social, and professional skills. They must be able to

quickly adapt to the world changes (Redhana, 2019) and also need to be creative, innovative, able to think critically, able to make correct decisions, and able to solve problems well. These abilities are parts of the teacher's higher-order thinking skills. In the bloom taxonomy, higher-order thinking skills (HOTS) are represented by the ability to analyze, evaluate, and create. Currently, it has been developed by a more recent theory by adding logic and reasoning indicators, problem-solving, and judgment. Therefore, teacher training is expected to be able to produce the best prospective teachers who possess these abilities.

The skills demanded in the 21st century are communication, collaboration, critical thinking, and problem-solving, as well as creativity and innovation (Arifin, 2017). Students can have it if the teacher can develop a well-planned lesson plan. But, the study of Haviz et al., (2020) said that the 21st-century skill of prospective teachers was low. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). The prospective teachers' 21st-century skills in science learning were found to be predicting each other (Zorlu & Zorlu, 2021). One approach that meets the purpose is scientific. The scientific approach aims to provide an understanding of gaining knowledge and understanding various materials using scientific procedures.

The scientific approach has the potential to promote HOTS by using scientific reasoning (Pradana, 2020). It consists of several main activities, namely observing, questioning, experimenting, associating, communicating, and networking (Pradana, 2020; Susantini et al., 2016). All of these scientific activities can potentially influence the HOTS. HOTS are mental processes that require students to manipulate information and ideas in a certain way that gives them new understanding and implications, for example combining ideas in the process of synthesizing, generalizing, explaining, and making hypotheses to conclude. It is related to cognitive abilities in analyzing, evaluating, and creating.

The success of research on HOTS in primary teacher education has not sufficiently addressed natural science learning, although the subject is essential to equip students with process skills. Natural science learning can empower 21st-century skills, especially HOTS through learning models, one of which is metacognition-based learning. Therefore, in science learning, it is recommended to apply various forms of learning that can optimally empower students' metacognitive skills (Fauzi & Sa'diyah, 2019). From the definition of natural science as a process, attitude, and product, it can be concluded that qualified natural science teachers have excellent thinking skills.

The success of the scientific approach and other approaches in the process of learning to teach has been accomplished. For example, the scientific approach which was modified with technology (Chang & Hwang, 2018; Hartman & Johnson, 2018; He et al., 2016; O'Flaherty & Phillips, 2015) and the modification of inquiry with collaboration models have been successfully achieved (Chebii et al., 2012; Kovanović et al., 2015; Mayordomo & Onrubia, 2015). This success is also accompanied by the measurement and development of HOTS instruments in learning. Among them were the success of analyzing HOTS on the 5th-grade social science multiple choice questions (Yuniar et al., 2019) and the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Ahmad et al. (2018) have successfully developed the HOTS instrument in Basic Mathematics subject in primary teacher education. However, this finding is limited to the assessment of mathematicians and linguists. Broader implementation needs to be done to test the instrument empirically. Ahmad et al. (2018) found that most students of primary teacher education have not demonstrated excellent skills in planning and implementing HOTS learning in primary schools. An instrument that has been tested, valid, and feasible based on experts' evaluation has been developed to measure the HOTS of primary teacher education students (60% of the students have poor HOTS) (Ahmad et al., 2018).

The above findings still have limitations in terms of substance and methodology. There is no valid question instrument that has been successfully developed to measure the students' HOTS of the elementary school education department in science learning. What is meant by valid here is that it has

been through testing by experts and empirically. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of the elementary school education department in science learning. This instrument can be used to see the students' HOTS so that the teacher training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

This study aims to develop a valid measurement tool in measuring the students' higher-order thinking skills of the elementary school education department. The designed product can be used in many similar institutions to analyze students' HOTS to be able to find weaknesses and solutions for improvement.

Methods

Research Design

This research and development study aims to produce HOTS instruments in the form of multiple-choice and essay questions. The final product was tested for measuring the quality through a content validation and empirical test. In this study, the 4D model by Thiagarajan, Semmel, and Semmel (Thiagarajan et al., 1974), which includes define, design, develop, and disseminate phases, was employed.

Define

This define phase is divided into three stages. The first stage is the initial objective analysis, the second is material analysis, and the last stage is the analysis of the learning outcomes, competence, and learning indicators, which are used to design the question indicators and items. This phase produces a list of materials that are considered complex by teacher students and used as material for developing this instrument (multiple choice and essay questions). Both question types were chosen because of their strengths in terms of effectiveness, ease of analysis, and practicality in measuring HOTS.

Design

The design phase produced more detailed product specifications which can be described as the following:

- the test questions consist of 10 multiple choice and 5 essay questions;
- each indicator (analysis, evaluation, creation) refers to Bloom taxonomy, consists of more than 2 questions.
- the instruments contain an introduction, guidelines, related materials, content outlines, question items, answer choices, answer sheets, and an answer key;
- the content outlines contain learning outcomes, learning indicators, problem indicators, cognitive level, number of question items, stimulus, answer keys, and scoring guidelines.

In addition to the question, the HMCEQ is also completed with a summary of the materials being tested to help students recall the materials. The results of the design phase are the first products that are ready to be tested by experts and prospective teachers.

Develop

At this phase, the initial product from the design phase is developed. This phase consists of content validity and constructs validity. The content validity involves experts at natural science, experts at evaluation studies, and experts at pedagogical in primary school. They were asked to

provide suggestions and assess the quality of HMCEQ. Specifically, the experts were asked to assess the instrument from the aspects of material selection, cognitive process category, the content of the test instrument, question type, question instruction, answer key, and language. The experts gave comments and suggestions on the question items as well as scores that indicate the quality using the assessment sheet. These experts' assessments were used to repair the instrument. The next process in the development stage is the empirical test. We involved 156 students of the primary teacher education department who are taking a natural science course to became the participants in the test. The test was used to determine validity, reliability, discrimination index, distractor efficiency, and difficulty index. The final product of the development phase is a valid HMCEQ that meets the experts' judgment and empirical testing. The HMCEQ is ready to be implemented in the dissemination stage.

Disseminate

The dissemination stage is in the form of product dissemination to the elementary school teacher education department association, especially the natural science lecturer. The dissemination was conducted online at a workshop of science curriculum review of the elementary school teacher education department. This dissemination aims to obtain input, corrections, suggestions, assessments, to improve the final product development so that it is ready for adoption by product users.

Participants

The research participants consist of subjects for testing and subjects for implementation. In the development step, 81 students in their 2nd year in primary teacher education were selected to participate. In contrast, in the dissemination step, 75 students in their 1st year who are taking a Natural Science course in the primary teacher education of Universitas Ahmad Dahlan, Yogyakarta, Indonesia took part in the research. Simple random sampling was used to select participants. Samples were taken randomly without considering the existing strata in the population (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

Instrument

Item Construction

The developed HMCEQ was designed based on natural science learning outcomes in primary teacher education. Two learning outcomes were elaborated into two learning indicators. These two learning indicators were expanded into ten problem indicators, which were represented by ten multiple-choice questions and five essay questions.

Experts' Appraisal

In addition to the test, the HMCEQ quality was also assessed by experts using the Delphi technique. The experts were asked to assess the aspects of HMCEQ in terms of material selection, cognitive process category, content of the test instrument, question type, question guidelines, and answer key, and language. The experts commented on the question items, made suggestions, and assessed the quality by giving a score in an assessment sheet. Experts' suggestions were used to revise the HMCEQ.

Data Analysis

The data obtained from the results of the validation test by experts and respondents were analyzed as a reference for product revision. The analysis was conducted during and after the data collection process. Qualitative analysis in this study was used to analyze data generated from experts' notes, comments, criticisms, and suggestions. The next step is the empirical test to determine validity, reliability, discrimination index, and difficulty index. The quality of the instrument (multiple choices and essay) were analyzed by Item Response Theory using the Rasch Model. The validity and reliability were tested to determine the quality of the considered questions based on the level of difficulty and the index of discrimination (Istiyono et al., 2020). The distractor efficiency of multiple-choice questions is obtained from the formula $DE = \frac{JPJ}{n}$. It is explained that DE= answer distribution for the particular option of an answer; JPJ = number of students who chose the option of an answer; and n = number of students. It can be said that the distractor functions if it is chosen by at least 5% of the testing participants (Hingorjo & Jaleel, 2012).

Findings

This research has succeeded in developing three HMCEQ sets to measure the students' higher-order thinking skills of the elementary school education department through the stages of define, design, development, and dissemination.

Define Phase

At the defined stage, the urgency of developing HMCEQ is based on the high need for HOTS measurement instruments for students of the elementary school education department. The instruments that have been used so far have not been adapted to HOTS-oriented learning outcomes. Although the learning process is required to empower HOTS, the facts on the ground show different things. Therefore, HMCEQ is a solution to solve this problem. Furthermore, an analysis of learning outcomes is carried out and the material of the human respiratory system was selected. This material was chosen because it is abstract and has high complexity. The results of the material analysis, including materials for study, course learning outcomes, and indicators of targeted competency, are presented in Table 1.

Table 1

Analysis of Learning Outcomes and Indicators

Materials	Course Learning Outcomes	Indicators
Organ Systems	Students can understand the motion system, digestive system, respiratory system, and blood circulatory system	Analyzing the structure and functions of the organs of the respiratory system
		Analyzing the respiratory problems experienced by people in the society

Design Phase

The design stage produced the instrument manual containing the test outline, test items (consisted of 10 multiple choice items and 5 essays), test direction, answer sheet, answer key, and scoring guide. At this stage, the blueprint for question items which is presented in Table 2 was designed.

Table 2

Examples of Blueprint for Question items to Measure HOTS

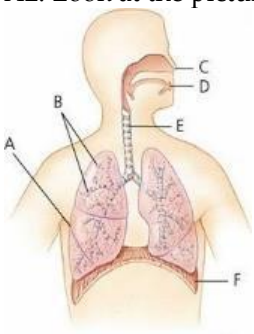
Learning Outcomes	Learning Indicators	Question Item Indicators	Number of Question Items	Stimulus	Cognitive Level
Students can understand the structure and functions of the organs of the respiratory system	Analyzing the structure and functions of the organs of the respiratory system	A statement is presented, students can confirm the anatomy and physiology of the lungs	A1 (Multiple choice)	Statement	C4
		An illustration is presented, students can confirm the exchange location between oxygen and carbon dioxide	A2 (Multiple choice)	Illustration	C5
		A story is presented, students can understand the right side sleeping	B2 (essay)	Story	C5

The guidelines above were formulated in the following questions.

Multiple Choice Questions

A1. The lungs function to transport oxygen from the air into the bloodstream.
It indicates that the lungs…
a. have a wide surface
b. have an elastic surface
c. are rich in capillary
d. are protected by a pleural membrane
e. have two lobes

A2. Look at the picture below!



In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide are indicated by letter…

a. A
b. B
c. C dan D
d. C dan D
e. E dan F

Essay Question

B2. Anton has a habit of sleeping on his right side. Right side sleeping is the best sleeping position that is beneficial for health, including the lungs. Explain the reasons!

Answer: ……

Each question has a different stimulus in the form of a statement, table, illustration, problem, experimental results, or statistical data. Each multiple-choice question has five answer choices (a, b, c, d, e), while the essay questions require a clear answer. For multiple-choice questions, each correct answer is given a score of 1, while the score for essay questions is 6. The scoring rubric for the above essay questions are:

0: didn't answer
2: answered but not related to the question
4: answered correctly but incomplete explanation
6: correct answer and full explanation

Development Phase

Validity Test

The development stage was conducted by developing the blueprint into question items, testing content validity, and conducting an empirical test. The content validity involves experts at natural science, experts at evaluation studies, and experts at pedagogical in primary school. Experts assessed the content validity regarding the aspects of the material, question guidelines, HOTS question type, question construction, question arrangement, answer key, and language use. The results of the experts' assessment can be seen in Table 3.

Table 3

Results of Product Assessment by Experts

Indeks	Validators	Value	Qualifications
1	Evaluation experts	79 %	Good
2	Pedagogical in primary school experts	83.3 %	Very Good
3	Natural science experts	81.3 %	Very Good
	Average	81.2 %	Very Good

The content validity shows an average value of 81.2%, which means that the validity was in a very good category. After the product was assessed by experts, it was tested again to measure the validity, reliability, discrimination index, distraction function, and difficulty index. On the test day, the students were given 30 minutes to read the material summary about the respiratory system. After that, the students were given 45 minutes to answer the questions. The results of the test item fit for multiple-choice items are described in Table 4.

Table 4

Validity Test Result of Multiple Choice Questions

Type of test	Item	X^2	Pr (> X^2)	Result
Multiple Choice	Item 1	22.9292	0.0003	fit
	Item 2	12.6841	0.0265	fit
	Item 3	5.9195	0.3141	misfit
	Item 4	22.6654	0.0004	fit

	Item 5	22.5403	0.0004	fit
	Item 6	9.2658	0.0989	misfit
	Item 7	28.5175	<0.0001	fit
	Item 8	16.6519	0.0052	fit
	Item 9	4.0696	0.5394	misfit
	Item 10	8.6818	0.1224	misfit

Note: Test items by model fit, p > 0.05: misfit

While the validity test for essay questions is described in Table 5.

Table 5

Validity Test Result Essay Questions

Item	R_{value}	Criteria
B1	0,548	Valid
B2	0.286	Invalid
B3	0,743	Valid
B4	0,203	Invalid
B5	0,470	Valid

Note. $R > R_{table}$ (0,367) = valid

Based on Tables 4 and 5, 3 items in multiple choices questions are a misfit and 7 items are fit, whereas 2 items in the essay questions are invalid. This can be caused by the difficulty index, distractor function, language, or terms in the question, as well as other factors related to the question construction. In this study, it is suspected that the cause of the two misfit multiple choices questions can be explained as the following.

Question B2: The stimulus for the question is very complex so that it did not help students much in analyzing the answer to the stimulus.

Question B4: It is too easy so that all students could answer the question correctly.

The follow-up activity that can be done is revising the two invalid questions. Therefore, the stimulus was adjusted for question B2, and the cognitive level for question B4 was increased by increasing the difficulty index. The revision process is shown in Table 6.

Table 6

Revision of Invalid Questions

Question	Before revision			After revision		
	Indicators	Questions	Cognitive Level	Indicators	Questions	Cognitive Level
B2	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration 682ea rit682 lungs inflated and deflated. In the process, there is a possibility that the lungs rub against the ribs or other organs. However, the lungs are not injured despite the friction. Why does this happen?	C5	A statement is presented, students can clarify the process of air exchange in the lungs	When we breathe, air exchange occurs in the lungs. In your opinion, how does the mechanism of air exchange in the lungs take place?	C5
B4	A problem is presented, students can identify the shortness of breath that happens in cold weather	Students were having a night gathering at Dieng plateau. Suddenly one of the students experienced shortness of breath because it was very cold and he could not 682ea rit. Why did it happen? What actions should be taken as the first aid to overcome shortness of breath?	C4	A problem is presented, students can predict the relation between carbon monoxide poisoning and respiratory system	Salsa's neighbor died yesterday. Based on the doctor's analysis, the cause of death was monoxide gas poisoning. Do you think carbon monoxide poisoning is related to the respiratory system?	C5

Reliability Test

The reliability test of HMCEQ is related to the accuracy of the test results (Heale & Twycross, 2015). Reliability is used to measure the consistency of a test. It is used to test the consistency of the question items when the test was taken repeatedly by the same object (Bajpai & Bajpai, 2014; Beck et al., 1994; Quaigrain & Arhin, 2017). The indicator that should be observed in the reliability values is a Kuder-Richardson 20 Test (KR-20). The KR-20 is suitable for determining the reliability coefficient of tests in which each item is parallel to the other. It is also suitable for questions that were scored by giving one point to the correct answers for each question, and no point to the wrong answers or unanswered questions (Sener & Tas, 2017). KR-20 test is useful for the internal consistency reliability of items. It is an equivalent measure for dichotomous items. Meanwhile, Cronbach's alpha test is an important and more useful test for the internal reliability of a questionnaire. It is a one-way concept of measuring the strength of that consistency (Singh, 2017). Based on the reliability test with KR-20 on multiple-choice questions, it resulted in a coefficient of 0.644 (reliable). Meanwhile, the reliability test using Cronbach's alpha in the essay questions resulted in a coefficient of 0.61 (reliable). This reliability value is sufficient and may be used for further research (Sumintono & Widhiarso, 2015).

Discrimination Indeks (DI) and Difficulty Indeks (DIF)

The discrimination index is the ability of a test item to distinguish between highly competent testing participants and those who are not (Panjaitan et al., 2018). The difficulty index is a measurement of the difficulty index of a question (Karelia et al., 2013). Analyzing the difficulty index of questions means classifying questions into easy, moderate, and difficult (Chauhan et al., 2015). The greater the item difficulty score, the more difficult the problem is, items are categorized as easy if they have a value of b nearly -2.00 logit, items are categorized as moderate if -1.00 logit <b <+1.00 logit and items are categorized as difficult if the value of b approaches +2.00 logit. Furthermore, items with a value of b> +2.00 logit into the very difficult category. In constructing test items, it should be noted that a balanced difficulty index should be used. The classification in the discriminant items is as follows. D ≥ 0.4 questions are very good, D between 0.3 - 0.39 questions are in the good category (questions are accepted without but need to be fixed), between 0.2 - 0.29 questions are sufficient / corrected, and D ≤ 0, 20 questions were discarded / bad questions (Vishnumolakala et al., 2016). The results of the difficulty index multiple choice questions showed in Figure 1 and the essay ones in Figure 2.

Figure 1

Result of Difficulty Index (DIF) of multiple-choice questions

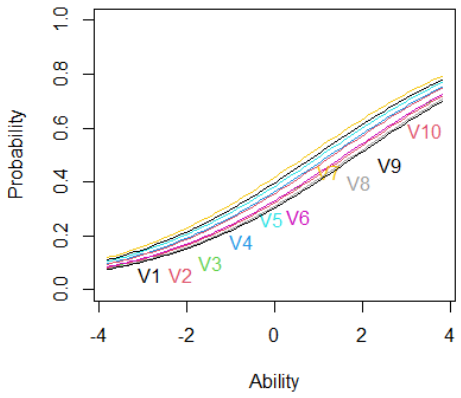


Figure 2

Result of Difficulty Index (DIF) of essay questions

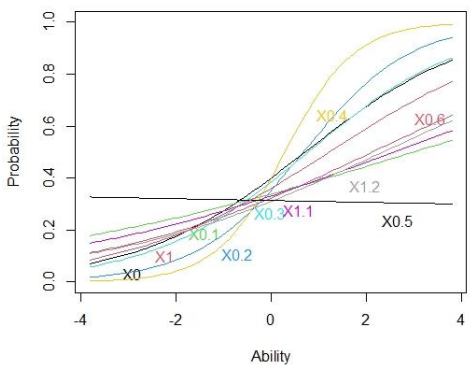


Figure 2 shows that the order of the difficulty index for multiple-choice questions from the easiest to the most difficult is V7-V1-V5-V4-V2-V6-V10-V8-V9, while for essay questions from easy to difficult are X1-X2-X4-X3-X5. The difficulty index and discriminant index data are shown in Table 7.

Table 7

Difficulty Index and Discriminant Index of Questions

Type of questions	Number	Difficulty index	Category	Discriminant Index	Category
Multiple Choice	1.	0.9376905	Moderate	0.530	Very good
	2.	1.6699686	Difficult	0.181	Discarded
	3.	1.6009005	Difficult	0.353	Good
	4.	1.2627958	Difficult	0.666	Very good
	5.	1.0665582	Moderate	0.618	Very good
	6.	1.5985782	Difficult	0.160	Discarded
	7.	0.7466097	Moderate	2.644	Very good
	8.	1.8095822	Difficult	0.093	Discarded
	9.	1.8804838	Difficult	0.067	Discarded
	10.	1.3292402	Difficult	0.315	Good
Essay	1.	-3,542	Easy	0.219	Sufficient
	2.	-2,631	Easy	0.843	Very good
	3.	2,331	Difficult	0.359	Good
	4.	1,491	Moderate	1.03	Very good
	5.	2,827	Difficult	0.313	Good

Distractor Efficiency (DE)

In multiple-choice questions, there is an option that functions as a distractor. The distractor works effectively if it is chosen by at least 5% of all testing participants (Hingorjo & Jaleel, 2012). The effectiveness of the distractor is how well the wrong option can deceive the testees who do not know the correct answer (Herrmann-Abell et al., 2011). The more testing participants were choosing the distractor, the more it functions appropriately. A good distractor will be chosen evenly by students who do not know the correct answer. On the contrary, a bad distractor will be chosen by an uneven number of students. Based on the analysis of the distractors, 26 distractors functioned effectively, and 14 distractors that did not function effectively. Because some distractors did not work properly, the answer choices were revised. An example of the revision process is presented in Table 8.

Table 8

Examples of Distractor Revision

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
A6	D	Influenza	Pneumonia	Bring answer choice closer to the answer key
	E	Lung cancer	Polyp	Bring answer choice closer to the answer key

HMCEQ, which had been declared feasible were used to analyze the HOTS of 79 prospective teachers taking the Natural Science course in the 5th semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

Figure 3

Analysis results of prospective teachers' HOTS

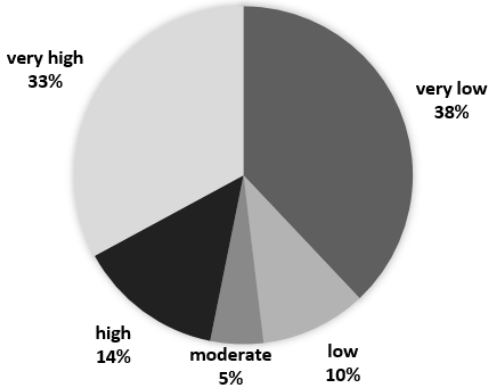


Figure 3 shows that most prospective teachers have very low HOTS (38%) and very high HOTS (33), while 14% have high HOTS, 10% low, and 5% moderate.

Disseminate Phase

The instrument has been complete in the dissemination phase through the Association of Elementary School Teacher Education Department in a lesson plan workshop.

Discussion

Evaluation is an activity of identifying, clarifying, and implementing criteria to achieve the success of a program (David et al., 2016). Evaluation can support the implementation of the curriculum, the certainty of school programs, the success of learning, and improve learning outcomes (Sugiyanta & Soenarto, 2016). A measuring instrument is required for evaluation. The instrument used must be valid and reliable in terms of content and construct because validity and reliability are important aspects of developing an instrument. An effective instrument can be used to obtain the required information (Tooth et al., 2013; Widodo & Sudarsono, 2016). Validity indicates what is supposed to be measured by the instrument (Azwar, 2015). The validity of the instrument can be identified from content validity and empirical test on each question item (Lissitz & Samuelsen, 2007). Therefore, content validity and empirical test were used in this study. The content validity is related to the rational analysis of the measured variables to determine the representation of the instrument with its ability to be measured (Retnawati, 2016). In measuring content validity for HMCEQ, natural science education experts and learning evaluation experts were involved.

In this study, the content validity obtained was 81.2% (closer to 100%). This means that the validity index agreement is higher than the items in the instrument, which are appropriate with the developed indicators. Additionally, it shows that the instrument has items that cover all variables that are intended to be measured. The content validity index can also be derived from empirical tests and experts' judgment (Creswell, 2012). Therefore, the empirical test is required to obtain more valid and reliable data. The summary in Table 5 shows that 2 items out of 45 items are invalid. It can be said that the test instrument has a high validity level.

Reliability shows that the consistency of an item is showing the same results when the test is conducted repeatedly (Eleje & Esomonu, 2018). The reliability test in the research showed Cronbach's

alpha coefficient of 0.605 for the multiple-choice questions; and 0.61 for the essay questions. It shows that the certainty of the consistency of the items in producing the same results repeatedly is within the percentage of 57% - 89%. The adequacy of an instrument is fulfilled when the instrument is reliable (Thaneerananon et al., 2016). Therefore, the multiple-choice question in this study was considered insufficient to meet the adequacy criteria, while the essay questions have met the criteria.

High validity indicates that the item or measuring instrument has truly measured the construct that is intended to be measured, while low reliability means that the measuring instrument is not able to produce a consistent value when measured in different situations. In predictive-criterion-related tests such as a test to measure higher-order thinking skills, validity is more important than reliability. When the validity value is satisfactory, the low-reliability value will not be a problem. In contrast, if the reliability is high and validity is low, it means that the instrument is proven to be able to produce consistent value in various situations, but has not been able to show the accurate measurement of a construct or something intended to be measured (Golafshani, 2003). Factors that affect the reliability index of a test are the number of items, construction of items, test instructions, test environment, scoring, and difficulty index (Jacobs & Chase, 1992; Postmes et al., 2013). To increase the reliability and validity of items, several alternatives can be taken, for example by selecting question items for the measuring instrument and testing the internal consistency and stability of the measuring instrument through a pilot study (Young et al., 2011). Other steps that can be taken include eliminating inter-observer measurement variations by involving trained and motivated people and eliminating intra-observer measurement variations by reducing sources of external variations such as boredom, fatigue, noisy environment, which affect research subjects and observers. Another alternative is to standardize the situation, context, or environment where the instrument is used (J. O. Chang et al., 2014; Ghosh et al., 2016; Postmes et al., 2013).

Difficulty index (DIF) describes the proportion of prospective teachers who answer an item correctly. It ranges from 0-1. The higher the proportion, the easier the item. The recommended difficulty range is from 0.3 - 0.7. Items that have DIF values <0.3 and> 0.7 are considered difficult and easy (Khoshaim & Rashid, 2016). DIF has a strong effect on variability in test scores (Thorndike et al., 1991). If the DIF is around 0.2-0.3 to 0.9, it can be concluded that the item is good and can be accepted. DIF is considered good when it is between 0.4 to 0.6. When DIF is less than 0.2, the item is too difficult and more than 0.9, it is too easy. It means that the item is unacceptable and needs modification or repair (Quaigrain & Arhin, 2017). In the product testing, the DIF obtained ranges from 0.1-0.9, indicating that the items are categorized as very easy to very difficult. Very easy items are placed at the beginning of the test as 'warm-up' questions. The aspects that make an item difficult include confusing language, distractors, problem stimulus, or even wrong answer keys (Hingorjo & Jaleel, 2012).

The quality of test items can be improved based on the actions taken in the analysis of distractor efficiency (DE), discrimination index (DI), and difficulty index (DIF). Some aspects that cause bad DI are the use of ambiguous language, neutral/doubtful answers, and wrong answer keys. Items showing DI must be reviewed again by content experts for revision to improve the standard of the test items. It is important to evaluate test items to find out the effectiveness in assessing students' knowledge based on DIF and DI (Karelia et al., 2013).

Distractor efficiency (DE) provides information about the overall quality of items (Burud et al., 2019). The selection of a good distractor can improve the test quality by affecting the difficulty index (Chauhan et al., 2015). However, further research on the effect of the number of distractors on the quality of the test still needs to be conducted. This study shows that out of 120 distractors, 49 distractors are categorized as non-functioning distractors (NFD). Multiple choice questions with more NFD indicate a high DIF compared to those with few NFD. The pattern of increasing DIF is 1NFD <2NFD <3NFD. However, multiple-choice questions with fewer NFD are not always difficult. The questions with a higher number of NFD are easier than those with a fewer number of NFD (Abdulghani et al., 2014).

This study provides useful findings that are valuable for the education sector because HMCEQ is a new instrument for measuring the HOTS of prospective primary school teachers. The implementation of various teacher training departments is strongly recommended so that the results of HOTS identification can provide an overview of prospective teachers' thinking skills. The teacher training department can prepare learning activities that can train and empower the prospective teachers' HOTS. This study has limitations in the scientific material used in the instrument is limited to the respiratory system. Therefore, it is necessary to develop instruments in other materials.

Conclusion and Implications

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of prospective teachers. The instrument consists of 10 multiple choice questions and five essay questions. Content validation shows a very good assessment result from experts. Based on the construct validity test, 7 questions are found to be fit, and three questions are misfits. The reliability test shows that the Cronbach Alpha Coefficient is 0,605 for the multiple-choice questions and 0.61 for the essay questions. Most items have a moderate and difficult difficulty index and a very good discrimination index. The test items that show a very good discrimination index tend to be difficult questions, and items that show a poor discrimination index tend to have varied difficulty indexes. The distractor efficiency shows that 59.2% of distractors worked well while the remaining 40.8% did not, which were revised based on the answer analysis of each item.

This valid instrument can be developed and implemented by elementary school teacher education for other courses to identify the HOTS of prospective teachers accurately. The results can reveal the weaknesses of prospective teachers' HOTS so that the institution can develop learning models that lead to the empowerment of HOTS.

Acknowledgements

This project was granted by the research institutions and community service Universitas Ahmad Dahlan, under the "Penelitian Dasar" Scheme Grant number PD-140/SP3/LPPM-UAD/2020.

References

Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnamperuma, G. (2014). The relationship between non-functioning distractors and item difficulty of multiple choice questions: A descriptive analysis. *Journal of Health Specialties*, *2*(4), 148. https://doi.org/10.4103/1658-600x.142784

Ahmad, S., Prahmana, R. C. I., Kenedi, A. K., Helsa, Y., Arianil, Y., & Zainil, M. (2018). The instruments of higher order thinking skills. *Journal of Physics: Conference Series*, *943*(1). https://doi.org/10.1088/1742-6596/943/1/012053

Ahmad, Syafri, Kenedi, A. K., & Masniladevi, M. (2018). Instrumen Hots Matematika Bagi Mahasiswa Pgsd. *JURNAL PAJAR (Pendidikan Dan Pengajaran)*, *2*(6), 905. https://doi.org/10.33578/pjr.v2i6.6530

Arifin, Z. (2017). Mengembangkan instrumen pengukur critical thinking skills siswa pada pembelajaran matematika abad 21. *Jurnal THEOREMS (The Original Research of Mathematics)*, *1*(2).

Azwar, S. (2015). *Tes Prestasi: fungsi dan pengembangan pengukuran prestasi belajar*. Pustaka Pelajar.

Bajpai, R., & Bajpai, S. (2014). Goodness of Measurement: Reliability and Validity. *International Journal of Medical Science and Public Health*, *3*(2), 112. https://doi.org/10.5455/ijmsph.2013.191120133

Beck, C. T., Keddy, B. A., & Cohen, M. Z. (1994). Reliability and Validity Issues in Phenomenological Research. *Western Journal of Nursing Research*, *16*(3), 254–267. https://doi.org/10.1177/019394599401600303

Bhakti, C. P., & Maryani, I. (2017). Peran LPTK dalam Pengembangan Kompetensi Pedagogik Calon

Guru. *Jurnal Pendidikan (Teori Dan Praktik)*, *1*(2), 98. https://doi.org/10.26740/jp.v1n2.p98-106

Burud, I., Nagandla, K., & Agarwal, P. (2019). Impact of distractors in item analysis of multiple choice questions. *International Journal of Research in Medical Sciences*, *7*(4), 1136-1139. DOI:10.18203/2320-60. https://doi.org/10.18203/2320-6012.ijrms20191313

Chang, J. O., Levy, S. S., Seay, S. W., & Goble, D. J. (2014). An Alternative to the Balance Error Scoring System. *Clinical Journal of Sport Medicine*, *24*(3), 256–262. https://doi.org/10.1097/JSM.0000000000000016

Chang, S. C., & Hwang, G. J. (2018). Impacts of an augmented reality-based flipped learning guiding approach on students' scientific project performance and perceptions. *Computers and Education*, *125*, 226–239. https://doi.org/10.1016/j.compedu.2018.06.007

Chauhan, P., Chauhan, G. R., Chauhan, B. R., Vaza, J. V., & Rathod, S. P. (2015). Relationship Between Difficulty Index and Distracter Effectiveness in Single Best-Answer Stem Type Multiple Choice Questions. *International Journal of Anatomy and Research*, *3*(4), 1607–1610. https://doi.org/10.16965/ijar.2015.299

Chebii, R., Wachanga, S., & Kiboss, J. (2012). Effects of Science Process Skills Mastery Learning Approach on Students' Acquisition of Selected Chemistry Practical Skills in School. *Creative Education*, *03*(08), 1291–1296. https://doi.org/10.4236/ce.2012.38188

Creswell, J. W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative* (4th ed.). Pearson Education Inc.

David, D., Kartowagiran, B., & Harjo, S. P. (2016). Evaluasi Dan Strategi Pengembangan SMA Indonesisch Nerderlandsche School (INS) Kayutanam. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(1), 27. https://doi.org/10.21831/pep.v20i1.7518

Eleje, L. I., & Esomonu, N. P. M. (2018). Test of Achievement in Quantitative Economics for Secondary Schools: Construction and Validation Using Item Response Theory. *Asian Journal of Education and Training*, *4*(1), 18–28. https://doi.org/10.20448/journal.522.2018.41.18.28

Fauzi, A., & Sa'diyah, W. (2019). Students' metacognitive skills from the viewpoint of answering biological questions: Is it already good? *Jurnal Pendidikan IPA Indonesia*, *8*(3), 317–327. https://doi.org/10.15294/jpii.v8i3.19457

Ghosh, S., Bowles, M., Ranmuthugala, D., & Brooks, B. (2016). Authentic assessment in seafarer education: using literature review to investigate its validity and reliability through rubrics. *WMU Journal of Maritime Affairs*, *15*(2), 317–336. https://doi.org/10.1007/s13437-015-0094-0

Golafshani, N. (2003). Understanding and Validity in Qualitative Research. *The Qualitative Report*, *8*(4), 597–607. https://doi.org/10.17763/haer.62.3.8323320856251826

Hartman, H., & Johnson, P. (2018). The effectiveness of multimedia for teaching drug mechanisms of action to undergraduate health students. *Computers and Education*, *125*, 202–211. https://doi.org/10.1016/j.compedu.2018.06.014

Haviz, M., Maris, I. M., Adripen, Lufri, David, & Fudholi, A. (2020). Assessing pre-service teachers' perception on 21st century skills in Indonesia. *Journal of Turkish Science Education*, *17*(3), 351–363. https://doi.org/10.36681/tused.2020.32

He, W., Holton, A., Farkas, G., & Warschauer, M. (2016). The effects of flipped instruction on out-of-class study time, exam performance, and student perceptions. *Learning and Instruction*, *45*, 61–71. https://doi.org/10.1016/j.learninstruc.2016.07.001

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. In *Evidence-Based Nursing* (Vol. 18, Issue 3, pp. 66–67). BMJ Publishing Group. https://doi.org/10.1136/eb-2015-102129

Herrmann-Abell, F, C., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*(2), 184–192.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of One-Best MCQs : the Difficulty Index , Discrimination Index and Distractor Efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Istiyono, E., Widihastuti, W., Supahar, S., & Hamdi, S. (2020). Measuring Creative Thinking Skills of Senior High School Male and Female Students in Physics (CTSP) Using the IRT-based PhysTCreTS. *Journal of Turkish Science Education*, *17*(4), 578–590. https://doi.org/10.36681/tused.2020.46

Jacobs, L. C., & Chase, C. I. (1992). *Developing and Using Tests Effectively. A Guide for Faculty*. Jossey-Bass Inc.

Karelia, B. N., Professor, A., Pillai, A., & Vegada, B. N. (2013). The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of Year II M.B.B.S students. *IeJSME*, *7*(2), 41–46.

Khoshaim, H. B., & Rashid, S. (2016). Assessment of the assessment tool: analysis of items in a non-MCQ mathematics exam. *International Journal of Instruction*, *9*(1), 119–132. https://doi.org/10.12973/iji.2016.9110a

Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *Internet and Higher Education*, *27*, 74–89. https://doi.org/10.1016/j.iheduc.2015.06.002

Lissitz, R. W., & Samuelsen, K. (2007). Further Clarification Regarding Validity and Education. *Educational Researcher*, *36*(8), 482–484. https://doi.org/10.3102/0013189x07311612

Mayordomo, R. M., & Onrubia, J. (2015). Work coordination and collaborative knowledge construction in a small group collaborative virtual task. *Internet and Higher Education*, *25*, 96–104. https://doi.org/10.1016/j.iheduc.2015.02.003

Nur Aini, D. F., & Sulistyani, N. (2019). Pengembangan Instrumen Penilaian E-Quiz (Electronic Quiz) Matematika Berbasis HOTS (Higher of Order Thinking Skills) untuk Kelas V Sekolah Dasar. *Edumaspul: Jurnal Pendidikan*, *3*(2), 1–10. https://doi.org/10.33487/edumaspul.v3i2.137

Nursalam, N., & Rasyid, M. R. (2016). Studi Kemampuan Mahasiswa Mendesain Perencanaan Pembelajaran Matematika di Sekolah Menengah Pertama Berbasis Pendekatan Saintifik. *MaPan: Jurnal Matematika Dan Pembelajaran*, *4*(1), 94–116. https://doi.org/10.24252/mapan.2016v4n1a8

O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *Internet and Higher Education*, *25*, 85–95. https://doi.org/10.1016/j.iheduc.2015.02.002

Panjaitan, R. L., Irawati, R., Sujana, A., Hanifah, N., & Djuanda, D. (2018). Item validity vs. item discrimination index: A redundancy? *Journal of Physics: Conference Series*, *983*(1). https://doi.org/10.1088/1742-6596/983/1/012101

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*(4), 597–617. https://doi.org/10.1111/bjso.12006

Pradana, V. (2020). Penggunaan pendekatan saintifik untuk meningkatkan kemampuan menyelesaikan soal hots pada materi karakteristik geografi Indonesia. *Didaktika Dwija Indria*, *8*(04). https://doi.org/10.20961/ddi.v8i04.39916

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186X.2017.1301013

Redhana, I. W. (2019). Mengembangkan Keterampilan Abad Ke-21 Dalam Pembelajaran Kimia. *Jurnal Inovasi Pendidikan Kimia*, *13*(1).

Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*, *2*(2), 155. https://doi.org/10.21831/reid.v2i2.11029

Sener, N., & Tas, E. (2017). Developing Achievement Test: A Research for Assessment of 5th Grade Biology Subject. *Journal of Education and Learning*, *6*(2). https://doi.org/10.5539/jel.v6n2p254

Singh, A. S. (2017). Common procedures for development, validity and reliability of a questionnaire. *International Journal of Economics, Commerce and Management*, *5*(5), 790–801. https://www.researchgate.net/profile/Mohamed_Hammad11/post/Reliability_and_Validity_of_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/20

17+common+procedures+for+development%2c+valıdıty+and+Reliability.pdf

Sugiyanta, S., & Soenarto, S. (2016). An evaluation model of educational quality assurance at junior high schools. *Research and Evaluation in Education*, *2*(2), 194. https://doi.org/10.21831/reid.v2i2.11118

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata.

Susantini, E., Faizah, U., Prastiwi, M. S., & Suryanti. (2016). Developing educational video to improve the use of scientific approach in cooperative learning. *Journal of Baltic Science Education*, *15*(6), 725–737.

Thaneerananon, T., Triampo, W., & Nokkaew, A. (2016). Development of a test to evaluate students' analytical thinking based on fact versus opinion differentiation. *International Journal of Instruction*, *9*(2), 123–138. https://doi.org/10.12973/iji.2016.929a

Thiagarajan, S., Semmel, M. ., & Semmel, D. . (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Leadership Training Institute/Special Education University of Minnesota. https://eric.ed.gov/?id=ED090725

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. Macmillan Publishing Co, Inc.

Tooth, J. A., Nielsen, S., & Armstrong, H. (2013). Coaching effectiveness survey instruments: Taking stock of measuring the immeasurable. *Coaching*, *6*(2), 137–151. https://doi.org/10.1080/17521882.2013.802365

Vishnumolakala, V. R., Southam, D. C., Treagust, D. F., & Mocerino, M. (2016). Latent constructs of the students' assessment of their learning gains instrument following instruction in stereochemistry. *Chemistry Education Research and Practice*, *17*(2), 309–319. https://doi.org/10.1039/c5rp00214a

Widodo, E., & Sudarsono, F. X. (2016). Developing an observation instrument for assessing the effectiveness of English teaching at vocational secondary schools. *Research and Evaluation in Education*, *2*(2), 135. https://doi.org/10.21831/reid.v2i2.8648

Young, D. L., Estocado, N., Landers, M. R., & Black, J. (2011). A pilot study providing evidence for the validity of a new tool to improve assignment of national pressure ulcer advisory panel stage to pressure ulcers. *Advances in Skin & Wound Care*, *24*(4), 168–175. https://doi.org/10.1097/01.asw.0000396304.90710.ea

Yuniar, M., Rakhmat, C. R., & Saepulrohman, A. (2019). Penggunaan Media Kartu Pecahan untuk Meningkatkan Pemahaman Siswa tentang Membandingkan Pecahan. *Penggunaan Media Kartu Pecahan Untuk Meningkatkan Pemahaman Siswa Tentang Membandingkan Pecahan*, *6*(1), 90–100.

Zorlu, Y., & Zorlu, F. (2021). Investigation of The Relationship Between Preservice Science Teachers' 21st Century Skills and Science Learning Self-Efficacy Beliefs with Structural Equation Model. *Journal of Turkish Science Education*, *18*(1), 1–16. https://doi.org/10.36681/tused.2021.49