

## BUKTI KORESPONDENSI

### ARTIKEL JURNAL INTERNASIONAL BEREPUTASI (Q2)

Judul Artikel	:	HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers
Nama Jurnal	:	Journal of Turkish Science Education
Penulis	:	<b>Ika Maryani</b> , Zuhdan Kun Prasetyo, Insih Wilujeng, Siwi Purwanti, Meita Fitriawanati

No	Nama Bukti	Tanggal Aktivitas
1.	Bukti Submit pertama	26 Juli 2020
2.	Bukti review dari reviewer 1 round 1	14 September 2020
3.	Bukti review dari reviewer 2 round 1	23 September 2020
4.	Bukti Revisi Artikel round 1	15 Desember 2020
5.	Bukti review dari reviewer 1 round 2	13 Januari 2021
6.	Bukti review dari reviewer 2 round 2	12 Februari 2021
7.	Bukti Revisi Artikel round 2	20 Maret 2021
8.	Bukti Accepted	31 Desember 2021
9.	Bukti Proofread	11 November 2021
10.	Bukti Copy editing	31 Desember 2021
11.	Bukti Publish	31 Desember 2021

← Back to Submissions



1018 / Ika Maryani et al. / HOTS Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills

[Library](#)

[Workflow](#) [Publication](#)

# BUKTI SUBMIT

[Submission](#) [Review](#) [Copyediting](#) [Production](#)

Submission Files				<a href="#">Q Search</a>
▶	 1866	<a href="#">ikamaryani, TUSED-1-without name.doc</a>	July 26, 2020	Article Text
▶	 1867	<a href="#">ikamaryani, TUSED-1-with name.doc</a>	July 26, 2020	Article Text
				<a href="#">Download All Files</a>

Pre-Review Discussions					<a href="#">Add discussion</a>
Name	From	Last Reply	Replies	Closed	
<i>No Items</i>					

99+ Compose

Mail

Chat

Spaces

Meet

Inbox 4,580

Starred

Snoozed

Important

Sent

Drafts 132

Categories

More

Labels

UAD

Q tused

UNIVERSITAS AHMAD DAHLAN

29 of 43

[tused] Editor Decision External Inbox x



İsa DEVECİ, Assoc. Prof. Dr., Kahramanmaras Sutcu Ima...

Sat, Sep 26, 2020, 6:47 PM

to me, Irma, Zuhdan, Insih, Siwi

Ika Maryani, Irma Rifda Syahada, Zuhdan Kun Prasetyo, Insih Wilujeng, Siwi Purwanti:

We have reached a decision regarding your submission to Journal of Turkish Science Education, "MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primary School Natural Science Teachers".

Our decision is to: Resubmit for Review

The reviewer comments are included at the bottom of this letter.

The reviewers pointed out some serious shortcomings and inadequacies in your work. If you think you can make these corrections, please revise your work. Please take the following points into consideration while sending your work to our journal for the second round.

1-When you revise your manuscript please highlight the changes you make in the manuscript by using the track changes mode in MS Word or by using bold or coloured text. (both manuscript with authors and blind manuscript)

2- As a separate file, upload a detailed author response form regarding reviewer' criticism or comments.

Once again, thank you for submitting your manuscript to Journal of Turkish Science Education and I look forward to receiving your revision.

Sincerely,

Dr İsa DEVECİ

Journal of Turkish Science Education Editorial Office

Reviewer(s)' Comments to Author:

Reviewer 1

First of all, congratulations on your effort, it seems to be a very comprehensive work. However, there are parts of your work that need technical revision. In addition, some points about your study need clarification.

Firstly;

- I find the abbreviation of the instrument incomplete, it does not cover high-order skills.

- Abstract needs to be rewritten, it is insufficient to cover the overall study. It is not clear how many questions were prepared for the instrument and to whom

**BUKTI  
PERMINTAAN  
RESUBMIT**

## **MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primary School Natural Science Teachers**

---

### **ABSTRACT**

This study aims to develop MCEQ for measure pre-service primary school natural science teachers' HOTS. This study used a 4-D research design. Evaluation experts, language experts, and natural science education experts were involved in content validation. The quality test conducted by experts showed that the average score of the question was 81.16 (very good). The validity test of the question set A and B demonstrated that all questions were valid. In contrast, in question set C, there were 13 questions classified as valid and two questions classified as invalid. The reliability showed fair, moderate, and high. The discrimination index showed low, moderate, high, and very high. The difficulty index showed very easy, easy, moderate, difficult, and very difficult. The distractor efficiency showed that 59.2% were functioning distractors, and 40.8% were non-functioning. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low.

**Keywords:** *MCEQ, instrument, validated, HOTS.*

---



## **INTRODUCTION**

Education in the 21st century requires students to have skills to learn and innovate, to use technology and information media, and to work and survive using life skills. Based on this change in the paradigm of learning in the 21st century, the LPTK (Institute of Teachers' Training) is required to produce qualified prospective teachers. (Bhakti & Maryani, 2017) explained that LPTK has a task to prepare professional teachers, educators for the nation's generation. Teachers are professional occupations that provide expert services and demand academic, pedagogical, social, and professional skills. Teachers are human resources in education who must be able to follow changes quickly (Redhana, 2019). Teachers must be creative, innovative, able to think critically, able to make decisions correctly, and able to solve problems well. These abilities are parts of the teacher's higher-order thinking skills. Therefore, LPTK is expected to be able to produce the best teacher candidates who possess these abilities.

Human resource skills that are demanded in the 21st century are communication, collaboration, critical thinking and problem solving, and creativity and innovation (Arifin, 2017). Students can possess these abilities if the teacher can develop well-planned learning plans. Learning plans that are designed must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rusydi Rasyid, 2016). One approach that meets the purpose is a scientific approach. The implementation of scientific approach in the 2013 curriculum in Indonesia was intended to provide an understanding of getting knowledge of and understanding various materials using scientific approach.

The scientific approach has the potential to maximize HOTS by using scientific reasoning (Pradana, 2020). The scientific approach consists of several main activities, namely observing, questioning, experimenting, associating, communicating, and networking (Pradana, 2020; Susantini, Faizah, Prastiwi, & Suryanti, 2016). All of these scientific activities can potentially influence higher-order thinking (HOTS). HOTS is a thought process that requires 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, making hypotheses to conclude. HOTS is related to cognitive abilities in analyzing, evaluating, and creating.

The success of 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 (S. C. Chang & Hwang, 2018; Hartman & Johnson, 2018; He, Holton, Farkas, & Warschauer, 2016; O'Flaherty & Phillips, 2015) and the modification of inquiry with collaboration models have been successfully achieved (Chebii, Wachanga, & Kiboss, 2012; Kovanović, Gašević, Joksimović, Hatala, & Adesope, 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, Rakhmat, & Saepulrohman, 2019), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). In addition, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. (Syafri 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 and is 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) (S. Ahmad et al., 2018). 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. 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. Therefore, this study aims to develop a valid MCEQ in measuring primary teacher education students' HOTS in natural science. The designed product can be used in many similar institutions to analyze students' HOTS to be able to find weaknesses and solutions for improvement.

## **AIMS**

This study aims to develop MCEQ (Multiple Choice and Essay Questions) for measure pre-service primary school natural science teachers' HOTS.

## **METHODS**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the pilot phase, 81 junior students in primary teacher education were selected to participate. In contrast, in the implementation phase, 75 freshmen who are taking a Natural Science course in the primary teacher education of Universitas Ahmad Dahlan took part in the research. Simple random sampling was used to select participants. The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development aim to produce a test instrument for 3 question sets in the form of multiple-choice and essay questions as the end product. The final product produced was then tested for measuring the quality through a process of validation and empirical test. In this study, the 4D model by Thiagarajan, Semmel, and Semmel, which includes define, design, develop, and disseminate was employed.

#### ***a. Define***

This define phase is divided into three stages. The first stage is the initial objective analysis. At this stage, the goal of developing a test instrument on natural science material based on higher-level thinking skills was set. The second stage is material analysis. The materials were identified based on the learning outcomes that must be achieved but are considered difficult by students. The third stage is the analysis of the learning outcomes, competence, and learning indicators, which are used to design the question indicators and items. Findings from the define process were used to write the question items in the form of multiple-choice and essay questions (MCEQ.) Both question types were chosen because of their strengths in terms of effectiveness, ease of analysis and practicality in measuring HOTS.

#### ***b. Design***

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

- the test questions consist of 3 question sets (A, B, C) and each set has 10 multiple choice and 5 essay questions;
- each HOTS indicator (analysis, evaluation, creation) consists of 3-6 questions that are evenly distributed in each question set;

- one question set contains 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 set, the MCEQ 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 participants.

*c. 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 science, experts at evaluation studies, and experts at pedagogical in primary school. The experts were chosen based on their expertise in the related field that corresponds to the product requirements. They were asked to provide suggestions and assess the quality of MCEQ. 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. This experts' assessment was used to repair the instrument. The next process in the develop stage is empirical test. Freshmen and junior students of primary teacher education who are taking a science course became the participants in the test. The test was used to determine validity, reliability, discrimination index, distractor efficiency, and difficulty index. All aspects must meet the high criteria. If any of the parameters receive a low score, it means that improvements will be made in accordance with the results of the item analysis. The final product of the develop phase is a valid MCEQ that meets the experts' judgment and empirical testing. The MCEQ is ready to be implemented in the disseminate stage.

*d. Disseminate*

This phase is the implementation of MCEQ, which has been developed for much wider areas, for example, for students or primary school natural science teachers in other areas. The purpose of this dissemination is to evaluate the effectiveness of MCEQ in measuring the HOTS of pre-service / in-service primary school natural science teachers.

## **Instrument**

### *a. Item Construction*

The developed MCEQ was designed based on natural science learning outcomes in primary teacher education. There were two learning outcomes that 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.

### *b. Experts' Judgment*

In addition to the test, the MCEQ quality was also assessed by experts using the Delphi technique. The experts were asked to assess the aspects of MCEQ 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 MCEQ.

## **Data Analysis**

The data obtained from the results of the validation test by experts and respondents were analyzed as a basis 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. Qualitative data analysis technique can be carried out by:

- collecting the data in the form of notes, comments, criticisms, and advice from experts which are obtained from the distribution of assessment questionnaires;
- collecting, selecting and classifying data based on test groups; and
- analyzing data and drawing conclusions from various results of the analysis to be used as a basis for taking action to improve the product being developed.

In quantitative data analysis, descriptive and inferential statistics are used. The experts' assessment of the quality of the MCEQ is based on the criteria described in Table 1.

*Table 1. MCEQ Quality Criteria Guidelines*

Categories	Test Results (scale 100)	Criteria	Action Taken
4	81-100	Very good	Implementation
3	61-80	Good	Implementation
2	41-60	Fair	Revision
1	< 41	Poor	Revision

If the score is  $\geq 60$  (good / very good), an empirical test to obtain construct validity will be carried out. The results from the test are analyzed to determine validity, reliability, discrimination index, difficulty index, and distractor efficiency.

### **Validity**

The validity of multiple-choice questions is obtained from the formula of point-biserial correlation  $r_{bis} = \frac{M_p - M_q}{S_t} \sqrt{pq}$ . The formula consists of  $r$  = point-biserial correlation

coefficient,  $M_p$  = number of respondents who answered correctly,  $M_q$  = number of respondents who answered incorrectly,  $S_t$  = standard deviation for all items,  $P$  = proportion of respondents who answered the question correctly, and  $Q$  = proportion of respondents who answered the question incorrectly. On the other hand, the validity of essay questions is obtained from product-moment correlation, formulated as

$$r_{xy} = \frac{N\sum x_1 y_1 - (\sum x_1)(\sum y_1)}{\sqrt{(N\sum x_1^2 - (\sum x_1)^2)(N\sum y_1^2 - (\sum y_1)^2)}}.$$

It is indicated that  $r_{xy}$ =correlation between  $x$  dan  $y$ ,  $x_1$ = value of the first  $x$ ,  $y_1$  =value of the first  $y$ , and  $N$  = number of value. A question is considered valid if the value of  $r$  that is calculated ( $r$  count) is greater ( $>$ ) than the value of  $r$  in the statistic table.

### **Reliability**

The reliability of multiple-choice questions is obtained from the KR-20 formula  $r_{KR-20} = \frac{k}{k-1} \left( \frac{1-\sum pq}{s^2} \right)$ . It is explained that  $r_{KR-20}$ = correlation coefficient with KR20;  $k$  = number of question items;  $p$ = proportion of correct answer on a particular item;  $q$ = proportion of incorrect answer on a particular item; and  $s^2$ = variance of the total score. On the other hand, the reliability of essay questions is obtained from the product-moment formula

$$r_{11} = \left( \frac{n}{n-1} \right) \left( \frac{1-\sum si^2}{\sum st^2} \right).$$

It is described that  $r_{11}$ = reliability coefficient of the test;  $n$ = number of question items;  $si^2$ = item variance; dan  $st^2$ = total variance. The criteria for reliability are as the following: 0.91–1.00 (very high); 0.71– 0.90 (high); 0.41– 0.70 (moderate); 0.21– 0.40 (low); and Negative – 0.20 (very low).

### ***Discrimination Index***

The discrimination index of multiple-choice questions is obtained from the formula  $DI = \frac{2(KA-KB)}{n}$ . It is described that DI = discrimination index; KA = number of students in the upper group who got the item correct; KB = number of students in the lower group who got the item correct; dan n = number of students. On the other hand, the discrimination index of essay questions is obtained from  $DI = \frac{Mean A - Mean B}{Skor\ maximum}$ . It is explained that DI = discrimination index; Mean A = mean of upper group students; Mean B = mean of lower group students; dan Skor maximum = maximum score of each item. The criteria for discrimination index are as the following: 0.71–1.00 (very different); 0.41–0.70 (different); 0.21–0.40 (fairly different); dan 0.00–0.20 (less different).

### ***Difficulty Index***

The difficulty index of multiple-choice questions is obtained from the formula  $DIF = \frac{JB}{n}$ . It is explained that DIF = difficulty index; JB = number of students who got the item correct; n = number of students. On the other hand, the difficulty index of essay questions is obtained from the formula  $DIF = \frac{Mean}{Maximum\ score}$ . It is described that DIF = difficulty index; Mean = mean of the score; Maximum score = maximum score of each item. The criteria for difficulty index are as the following: 0.71–1.00 (easy); 0.31–0.70 (moderate); and 0.00–0.30 (difficult).

### ***Distractor Efficiency***

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 distractor functions if it is chosen by at least 5% of the testee (Hingorjo & Jaleel, 2012).

## **FINDINGS**

This research has succeeded in developing three MCEQ sets to measure the HOTS of pre-service primary school natural science teachers through the stages of define, design, develop, and dissemination. At the define stage, the urgency of developing MCEQ is based on the high need for HOTS measurement instruments for students. 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, MCEQ is a solution to solve this problem. Furthermore, an analysis of learning outcomes is carried out and the material of the human respiratory system were 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 2.

*Table 2. Analysis of Learning Outcomes and Indicators*

<b>Materials</b>	<b>Course Learning Outcomes</b>	<b>Indicators</b>
<b>Organ Systems</b>	Students are able to 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

The next step after the define stage is the develop stage. At this stage, the blueprint for question items which is presented in Table 3 was designed.

*Table 3. Examples of Blueprint for Question items to Measure HOTS*

<b>Learning Outcomes</b>	<b>Learning Indicators</b>	<b>Question Item Indicators</b>	<b>Number of Question Items</b>	<b>Stimulus</b>	<b>HOTS Level</b>
Students are able to understand the structure and functions of the organs of the respiratory system	Analysing 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
<b>Etc...</b>					



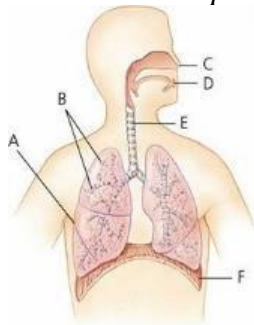
The guidelines above were formulated in the following questions.

*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 pleural membrane*
- e. have two lobes*

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

- 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 lungs. Explain the reasons!*

*Answer: .....*

Each question has a different stimulus in the form of the 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 clear answers from the students. For multiple-choice questions, each correct answer is given a score of 1, while the score for essay question is 6.

### Validity

The development stage was conducted by developing the guidelines into question items, testing content validity, and conducting an empirical test on the product. The content validity test involved experts in natural science education, learning evaluation, and language. 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 4.

*Table 4. Results of Product Assessment by Experts*

No	Validators	Value	Qualifications
1	Evaluation experts	79 %	Good
2	Language experts	83.3 %	Very Good
3	Natural science education experts	81.3 %	Very Good
	<b>Average</b>	81.2 %	Very Good

The content validity shows an average value of 81.2%, meaning 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 test results are described in Table 5.

*Table 5. Validity Test Result*

Question	Package A		Package B		Package C	
	$R_{value}$	Criteria	$R_{value}$	Criteria	$R_{value}$	Criteria
A1	0,550	Valid	0,445	Valid	0,443	Valid
A2	0,498	Valid	0,510	Valid	0,474	Valid
A3	0,487	Valid	0,617	Valid	0,403	Valid
A4	0,511	Valid	0,442	Valid	0,426	Valid
A5	0,476	Valid	0,730	Valid	0,599	Valid
A6	0,431	Valid	0,401	Valid	0,497	Valid
A7	0,387	Valid	0,474	Valid	0,500	Valid
A8	0,387	Valid	0,570	Valid	0,409	Valid
A9	0,397	Valid	0,401	Valid	0,705	Valid
A10	0,479	Valid	0,467	Valid	0,416	Valid

Question	Package A		Package B		Package C	
	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria
	N: 29, R <sub>table</sub> : 0,367		N: 25, R <sub>table</sub> : 0,367		N: 27, R <sub>table</sub> : 0,367	
<b>B1</b>	0,693	Valid	0,785	Valid	0,548	Valid
<b>B2</b>	0,608	Valid	0,517	Valid	0,286	<b>Invalid</b>
<b>B3</b>	0,746	Valid	0,474	Valid	0,743	Valid
<b>B4</b>	0,796	Valid	0,471	Valid	0,203	<b>Invalid</b>
<b>B5</b>	0,900	Valid	0,794	Valid	0,470	Valid

Based on Table 5, all items in question set A and B are valid, but in question set C, two items are invalid. A question is said to be valid if it measures what it is intended to measure. An invalid test produces data that is irrelevant to the measurement objective. This can be caused by the difficulty index of the question, distractor function, use of 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 invalid essay 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 analysing 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	A problem is presented, students can identify the shortness of breath 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 bear it. 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**

The reliability test of MCEQ 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 (Beck, Keddy, & Cohen, 1994). The test is said to be reliable or consistent if the scores are similar when the test is taken several times. This research used two methods for the reliability test.

- a) *Kuder-Richardson 20* is a special form of *Cronbach's alpha*. The value ranges from 0-1, with value closes to 1 indicating reliability. This method is used to find the internal consistency coefficient of multiple-choice questions (Quaigrain & Arhin, 2017).
- b) *Cronbach's alpha* is a method that will be used in analyzing essay questions. It is a coefficient of internal consistency and is widely used in social sciences, business, nursing, and other disciplines. It is the average of all split-half reliability estimates of an instrument and is usually used to estimate the reliability of psychometric tests for a sample of testees (Bajpai & Bajpai, 2014).

The results of the reliability test on the three question sets are presented in Table 7.

*Table 7. Results of Questions Reliability Analysis*

<b>Types of Question</b>	<b>Set A Question</b>		<b>Set B Question</b>		<b>Set C Question</b>	
	<b>R<sub>value</sub></b>	<b>Criteria</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
Multiple Choice	0.57	fair	0.67	moderate	0.65	moderate
Essay	0.89	high	0.70	moderate	0.81	high

If the reliability coefficient is within the range of 0.81 to 1.0, it indicates high reliability, 0.61-0.80 indicates moderate reliability, 0.41-0.60 indicates fair reliability, 0.10-0.40 indicates low reliability, and <0.10 indicates that the question is unreliable (Golafshani, 2003).

## **Discrimination Index (DI)**

Discrimination index is the ability of a test item to distinguish between highly competent testees and those who are not (Panjaitan, Irawati, Sujana, Hanifah, & Djuanda, 2018). Different methods to analyze the discrimination index of objective questions and essay questions were employed. Figure 1 shows the measurement results of the discrimination index for set A, B, and C questions.

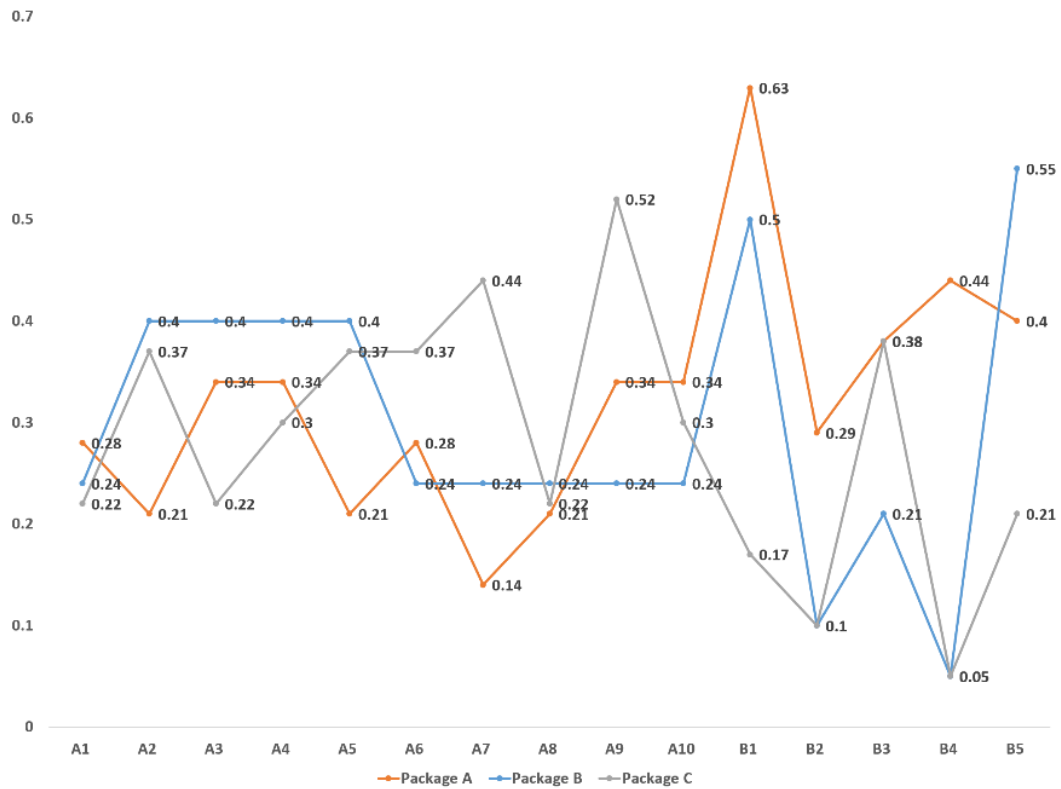


Figure 1. Result of the discrimination index

The data in Figure 1 were categorized based on the ability of the question items to distinguish testee. From these results, the question item is said to have a very good discrimination index if it has a DI of 0.71-1.00; good discrimination index if it has a DI of 0.41-0.70; sufficient discrimination index if it has a DI of 0.21-0.40; poor discrimination index if it has a DI of 0.00-0.20; and if the DI is negative, all question items are said to be bad [26]. The categorization of the Discrimination Index of the question items is shown in Table 8.

Table 8. The categorization of discrimination index

Package	Discrimination index			
	poor	sufficient	good	very good
A	A7, A9	A1, A3, A4, A5, A6, A8, B1, B2, B3, B4	A2, A10	-
B	A4, A6, A8, B2, B3, B4	A1, A2, A7, A9, B1, B5	A5, A10	A3
C	A2, A10, B1, B2, B3, B4, B5	A4, A8, A9	A1, A3, A5, A6, A7	-

### Difficulty Index (DIF )

Difficulty index is a measurement of the difficulty index of a question (Karelia, Professor, Pillai, & Vegada, 2013). Analyzing the difficulty index of questions means classifying questions into easy, moderate, and difficult (Chauhan, Chauhan, Chauhan, Vaza, & Rathod, 2015). In constructing test items, it should be noted that a balanced difficulty index should be used. The results of the difficulty index measurement are presented in Figure 2.

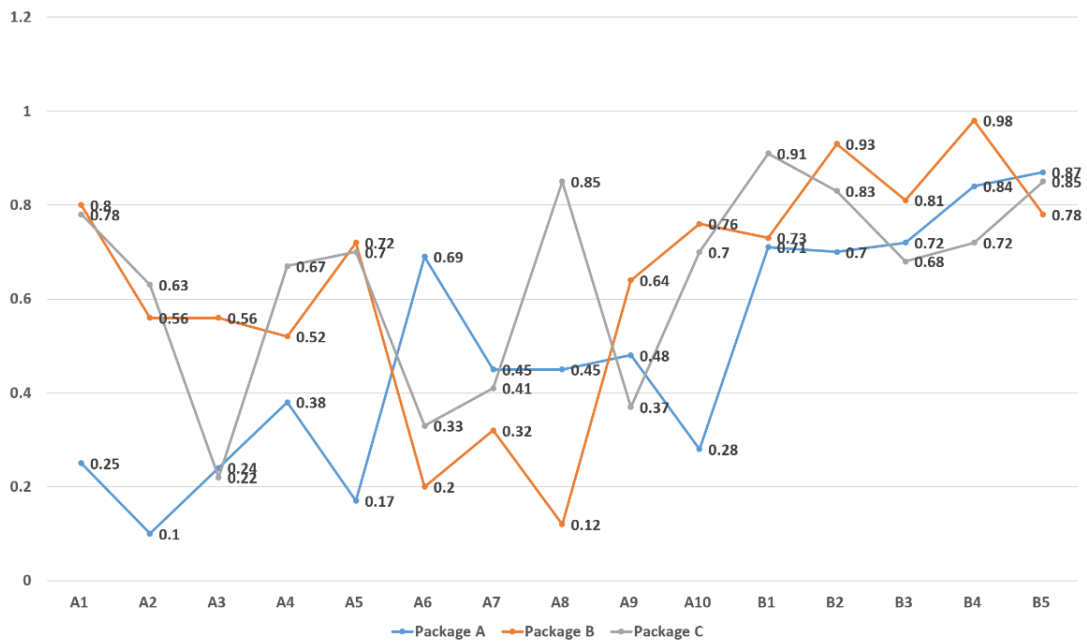


Figure 2. Result of Diffilucty Index (DIF)

A question item is called easy if many testees answer it correctly (DIF: 0.71-1.00). It is called sufficient if there is a balanced number of testees who answer it correctly and incorrectly (DIF: 0.31-0.70). It is called difficult if few testees answer it correctly (DIF: 0.00-0.30). In addition, it is called too easy if the value of P is equivalent to 1.00. An appropriate test item generally has P-value that ranges from 0.15 to 0.85 (Brown & Hudson, 2002). Based on the criteria and figure 2, the question set A has the following proportion of questions: 20% (difficult), 20% (moderate), and 60% (easy). The proportion in question set B shows that 13.33% of questions are classified as difficult, 20% are moderate, and 66.67% is easy. In question set C, 13.33% of questions are classified as difficult, 20% are moderate, and 66.67% is easy.

### *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 testees (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, F, & DeBoer, 2011).

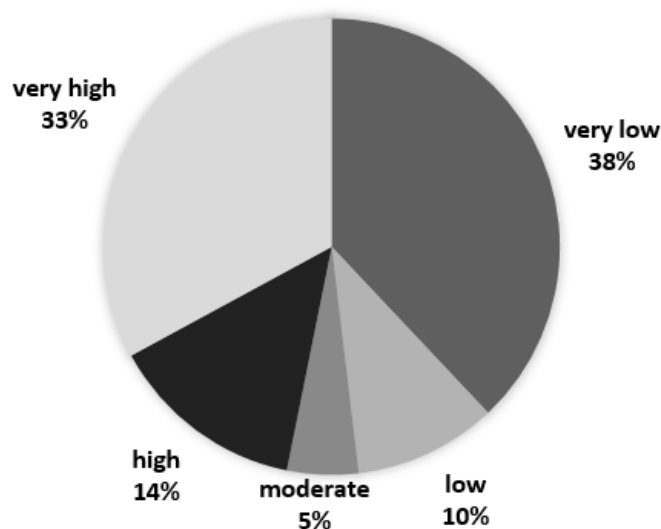
The more testees were choosing the distractor, the more it functions appropriately. A good distractor will be chosen evenly by students who do not know the correct answer. On the contrary, a bad distractor will be chosen by an uneven number of students. Based on the analysis of the distractors, the question set A contains 26 distractors that functioned effectively and 14 distractors that did not function effectively. Similarly, question set B shows that 26 distractors functioned effectively, and 14 distractors did not function effectively. Question set C shows that 19 distractors functioned effectively, and 21 distractors 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 9.

*Table 9. Example of Distractor Revision*

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
A6	D	Influenza	Pneumonia	Bring answer choice closer to the answer key
	E	Lung cancer	Polyp	Bring answer choice closer to the answer key
<b>Question set B</b>				
A4	E	Exchange of oxygen in the nasal cavity with CO <sub>2</sub> in the tissues	Exchange of oxygen in the bronchial cavity with CO <sub>2</sub> in the tissues	Bring answer choice closer to the answer key
<b>Question set C</b>				
A3	B	The process of inspiration in the lungs	The process of managing oxygen in the lungs	Bring answer choice closer to the answer key



MCEQ, which had been declared feasible were used to analyze the HOTS of 79 students taking the Primary School Natural Science Learning Development course in the 5<sup>th</sup> semester of primary teacher education. The results of the analysis are presented in Figure 3.



*Figure 3. Analysis results of students' HOTS*

## **DISCUSSION**

Evaluation is an activity of identifying, clarifying, and implementing criteria to achieve the success of a program (David, Kartowagiran, & Harjo, 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, Nielsen, & Armstrong, 2013; Widodo & Sudarsono, 2016). Validity indicates what is supposed to be measured by the instrument (Azwar, 2002). 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 MCEQ, natural science education experts and learning evaluation experts were involved.

In this study, the content validity obtained was 81.2% (closer to 1). 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 0.57; 0.65; 0.67; 0.70; 0.81; and 0.89. 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's reliability is higher than 0.70 (Thaneerananon, Triampo, & Nokkaew, 2016). Therefore, the three multiple-choice question sets in this study are considered insufficient to meet the adequacy criteria, while the three 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, Haslam, & Jans, 2013). In order to increase the reliability and validity of items, a number of 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, Estocado, Landers, & Black, 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/environment where the instrument is used (J. O. Chang,

Levy, Seay, & Goble, 2014; Ghosh, Bowles, Ranmuthugala, & Brooks, 2016; Postmes et al., 2013).

Difficulty index (DIF) describes the proportion of students 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 value  $<0.3$  and  $>0.7$  are considered difficult and easy (Khoshaim & Rashid, 2016). DIF has a strong effect on variability in test scores (Thorndike, Cunningham, Thorndike, & Hagen, 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 test, 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 key (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 key. 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, Nagandla, & Agarwal, 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  $1\text{NFD} < 2\text{NFD} < 3\text{NFD}$ . 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, Ahmad, Aldrees, Khalil, & Ponnampereuma, 2014).

This study provides useful findings that are valuable for the education sector because MCEQ 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 are able to provide an overview of students' thinking skills. The teacher training department can prepare learning activities that can train and empower the prospective teachers' HOTS.

## **CONCLUSION**

This research has succeeded in producing three sets of MCEQ on natural science to measure the higher-order thinking skills of students. Each question set consists of 10 multiple choice questions and five essay questions. Experts' validation shows a very good assessment result. Based on the construct validity test, 43 questions are found to be valid, and two questions are invalid. These invalid questions have been revised based on the item analysis. The reliability test shows that the criteria are sufficient, high, and very high, with Rvalue between 0.57 - 0.89. Most items have an accept difficulty index and a very good discrimination index. The discrimination index is in the moderate to very high category, while the difficulty index is in the category of very easy, medium, difficult, and very difficult. Revision, particularly on very difficult and very easy questions, was done. The test items that show a very good discrimination index tend to be difficult questions, and items that show poor discrimination index tend to have varied difficulty index. The distractor efficiency shows that 59.2% of distractors are functioning distractors while the remaining 40.8% are non-functioning distractors, which were revised based on the answer analysis of each item. This valid instrument can be developed and implemented by primary teacher education for other courses to identify the HOTS of prospective teachers accurately. The results can reveal the weaknesses of pre-service 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.

## **REFERENCE**

Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *Tes Prestasi: fungsi dan pengembangan pengukuran prestasi belajar*. Yogyakarta: 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>
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- 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. <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.). Boston-USA: 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>
- 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>
- 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. *Evidence-Based Nursing*, 18(3), 66–67. <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.
- Jacobs, L. C., & Chase, C. I. (1992). *Developing and Using Tests Effectively. A Guide for Faculty*. San Francisco, CA: 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. Retrieved from <https://eric.ed.gov/?id=EJ1086950>
- 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., & Rusydi Rasyid, M. (2016). Studi Kemampuan Mahasiswa Mendesain Perencanaan Pembelajaran Matematika Di Sekolah Menengah Pertama Berbasis Pendekatan Saintifik. *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>
- 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>
- 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>
- Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. Macmillan: 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>

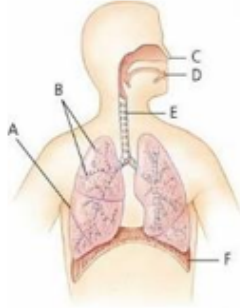


- Widodo, E., & Sudarsono, F. X. (2016). Developing an observation instrument for assessing the effectiveness of English teaching at vocational secondary schools. *Research and Evaluation in Education*, 2(2), 135. <https://doi.org/10.21831/reid.v2i2.8648>
- Young, D. L., Estocado, N., Landers, M. R., & 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.

## APPENDIXES

### A sample item for the Multiple Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# BUKTI PROSES REVIEW ROUND 1

Journal of Turkish Science Education


[← Back to Submissions](#)

Submission

Review

Copyediting

Production

Round 1

Round 2

## Round 1 Status

The submission must be resubmitted for another review round.

## Notifications

<a href="#">[tused] Editor Decision</a>	2020-09-26 11:46 AM
<a href="#">[tused] Editor Decision</a>	2020-09-26 11:48 AM
<a href="#">[tused] Editor Decision</a>	2021-02-17 08:47 AM
<a href="#">[tused] Editor Decision</a>	2021-04-06 09:29 AM
<a href="#">[tused] Editor Decision</a>	2021-12-31 11:23 AM

## Reviewer's Attachments

2046	Reviewer 1	September 14, 2020
2083	Reviewer 2	September 23, 2020

## Revisions



▶  2535	Article Text, Ika Maryani-revision-without name.docx	December 15, 2020	Article Text
▶  2536	Article Text, revision-with name.docx	December 15, 2020	Article Text

## Review Discussions

Name	From	Last Reply	Replies	Closed
<a href="#">Status of your manuscript</a>	ideveci 2020-12-01 08:47 AM	ikamaryani 2020-12-15 02:15 AM	2	<input type="checkbox"/>
▶ <a href="#">information</a>	ikamaryani 2021-01-20 07:04 AM	-	0	<input type="checkbox"/>
▶ <a href="#">article revision</a>	ikamaryani 2021-03-16 07:00 AM	ideveci 2021-03-27 10:54 AM	3	<input type="checkbox"/>

Name	From	Last Reply	Replies	Closed
<a href="#">▶ article revision</a>	ikamaryani 2021-03-29 07:17 AM	ikamaryani 2021-04-01 09:39 AM	4	<input type="checkbox"/>



Q tused



99+

Compose

Mail

Inbox

4,580

Chat

Starred

Snoozed

Spaces

Important

Sent

Meet

Drafts

132

Categories

More

Labels

UAD

29 of 43

[tused] Editor Decision External Inbox x



İsa DEVECİ, Assoc. Prof. Dr., Kahramanmaras Sutcu Ima...

Sat, Sep 26, 2020, 6:47 PM

to me, Irma, Zuhdan, Insih, Siwi

Ika Maryani, Irma Rifda Syahada, Zuhdan Kun Prasetyo, Insih Wilujeng, Siwi Purwanti:

We have reached a decision regarding your submission to Journal of Turkish Science Education, "MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primar School Natural Science Teachers".

Our decision is to: Resubmit for Review

The reviewer comments are included at the bottom of this letter.

The reviewers pointed out some serious shortcomings and inadequacies in your work. If you think you can make these corrections, please revise your work. Please take the following points into consideration while sending your work to our journal for the second round.

1-When you revise your manuscript please highlight the changes you make in the manuscript by using the track changes mode in MS Word or by using bold or coloured text. (both manuscript with authors and blind manuscript)

2- As a separate file, upload a detailed author response form regarding reviewer' criticism or comments.

Once again, thank you for submitting your manuscript to Journal of Turkish Science Education and I look forward to receiving your revision.

Sincerely,

Dr İsa DEVECİ

Journal of Turkish Science Education Editorial Office

Reviewer(s)' Comments to Author:

Reviewer 1

First of all, congratulations on your effort, it seems to be a very comprehensive work. However, there are parts of your work that need technical revision. In addition, some points about your study need clarification.

Firstly;

- I find the abbreviation of the instrument incomplete, it does not cover high-order skills.

- Abstract needs to be rewritten, it is insufficient to cover the overall study. It is not clear how many questions were prepared for the instrument and to whom

EMAIL  
PROSES  
REVIEW ROUND 1

# HASIL REVIEW DARI REVIEWER 1 ROUND 1

## **MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primary School Natural Science Teachers**

**Commented [A1]:** The abbreviation of the name of your instrument cannot fully explain what it is measuring. Multiple choice and essay questions for what? Higher order thinking skills?

### **ABSTRACT**

This study aims to develop MCEQ ~~for to~~ measure pre-service primary school natural science teachers' HOTS. This study used a 4-D research design. Evaluation experts, language experts, and natural science education experts were involved in content validation. The quality test conducted by experts showed that the average score of the question was 81.16 (very good). The validity test of the question set A and B demonstrated that all questions were valid. In contrast, in question set C, there were 13 questions classified as valid, and two questions ~~were~~ classified as invalid. ~~The reliability showed fair, moderate, and high.~~ The discrimination index showed low, moderate, high, and very high. The difficulty index showed very easy, easy, moderate, difficult, and very difficult. The distractor efficiency showed that 59.2% were functioning distractors, and 40.8% were non-functioning. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low.

**Commented [A2]:** The abstract is insufficient to cover the overall study. It is not clear how many questions were prepared for the instrument and to whom.

**Commented [A3]:** What is the meaning of this sentence? Have you calculated any reliability scores?

**Keywords:** *MCEQ, instrument, ~~validated, HOTS,~~ primary school natural science teachers*

**Commented [A4]:** please write open

**Formatted:** Strikethrough

**Formatted:** Turkish (Türkiye)

## INTRODUCTION

Education in the 21st century requires students to have skills to learn and innovate, to use technology and information media, and to work and survive using life skills. Based on this change in the paradigm of learning in the 21st century, the LPTK (Institute of Teachers' Training) is required to produce qualified prospective teachers. (Bhakti & Maryani, 2017) explained that LPTK has a task to prepare professional teachers, educators for the nation's generation. Teachers are professional occupations that provide expert services and demand academic, pedagogical, social, and professional skills. Teachers are human resources in education who must be able to follow changes quickly (Redhana, 2019). Teachers must be creative, innovative, able to think critically, able to make decisions correctly, and able to solve problems well. These abilities are parts of the teacher's higher-order thinking skills. Therefore, LPTK is expected to be able to produce the best teacher candidates who possess these abilities.

Human resource skills that are demanded in the 21st century are communication, collaboration, critical thinking and problem solving, and creativity and innovation (Arifin, 2017). Students can possess these abilities if the teacher can develop well-planned learning plans. Learning plans that are designed must be adjusted to the demands of the curriculum and must allow students to think and ~~analyze~~analyse critically (Nursalam & Rusydi Rasyid, 2016). One approach that meets the purpose is a scientific approach. The implementation of the scientific approach in the 2013 curriculum in Indonesia was intended to provide an understanding of getting knowledge of and understanding various materials using the scientific approach.

The scientific approach has the potential to maximize HOTS by using scientific reasoning (Pradana, 2020). The scientific approach consists of several main activities, namely observing, questioning, experimenting, associating, communicating, and networking (Pradana, 2020; Susantini, Faizah, Prastiwi, & Suryanti, 2016). All of these scientific activities can potentially influence higher-order thinking (HOTS). HOTS is a thought process that requires 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, making hypotheses to conclude. HOTS is related to cognitive abilities in analyzing, evaluating, and creating.

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 (S. C. Chang & Hwang, 2018; Hartman & Johnson, 2018; He, Holton, Farkas, & Warschauer, 2016; O'Flaherty & Phillips, 2015) and the modification of inquiry with collaboration models have been successfully achieved (Chebii, Wachanga, & Kiboss, 2012; Kovanović, Gašević, Joksimović, Hatala, & Adesope, 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, Rakhmat, & Saepulrohman, 2019), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). In addition, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. (Syafri 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 and is 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) (S. Ahmad et al., 2018). 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. 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. Therefore, this study aims to develop a valid MCEQ in measuring primary teacher education students' HOTS in natural science. The designed product can be used in many similar institutions to analyze students' HOTS to be able to find weaknesses and solutions for improvement.

#### **AIMS**

This study aims to develop MCEQ (Multiple Choice and Essay Questions) ~~for to~~ measure pre-service primary school natural science teachers' HOTS.



## **METHODS**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the pilot phase, 81 junior students in primary teacher education were selected to participate. In contrast, in the implementation phase, 75 freshmen who are taking a Natural Science course in the primary teacher education of Universitas Ahmad Dahlan took part in the research. Simple random sampling was used to select participants. The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development aim to produce a test instrument for 3 question sets in the form of multiple-choice and essay questions as the end product. The final product produced was then tested for measuring the quality through a process of validation and empirical test. In this study, the 4D model by Thiagarajan, Semmel, and Semmel ([Reference?](#)), which includes define, design, develop, and disseminate was employed.

#### *a. Define*

This define phase is divided into three stages. The first stage is the initial objective analysis. At this stage, the goal of developing a test instrument on natural science material based on higher-level thinking skills was set. The second stage is material analysis. The materials were identified based on the learning outcomes that must be achieved but are considered difficult by students. The third stage is the analysis of the learning outcomes, competence, and learning indicators, which are used to design the question indicators and items. Findings from the define process were used to write the question items in the form of multiple-choice and essay questions (MCEQ.) Both question types were chosen because of their ~~strenghts~~ strengths in terms of effectiveness, ease of analysis, and practicality in measuring HOTS.

#### *b. Design*

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

- the test questions consist of 3 question sets (A, B, C) and each set has 10 multiple choice and 5 essay questions;
- each HOTS indicator (analysis, evaluation, creation) consists of 3-6 questions that are evenly distributed in each question set;

- one question set contains 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 set, the MCEQ 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 participants.

*c. 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 science, experts at evaluation studies, and experts at pedagogical in primary school. The experts were chosen based on their expertise in the related field that corresponds to the product requirements. They were asked to provide suggestions and assess the quality of MCEQ. 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~~ ~~This~~ experts' assessments ~~were~~ used to repair the instrument. The next process in the develop stage is the empirical test. Freshmen and junior students of primary teacher education who are taking a science course became the participants in the test. The test was used to determine validity, reliability, discrimination index, distractor efficiency, and difficulty index. All aspects must meet the high criteria. If any of the parameters receive a low score, it means that improvements will be made in accordance with the results of the item analysis. The final product of the develop phase is a valid MCEQ that meets the experts' judgment and empirical testing. The MCEQ is ready to be implemented in the disseminate stage.

*d. Disseminate*

This phase is the implementation of MCEQ, which has been developed for much wider areas, for example, for students or primary school natural science teachers in other areas. The purpose of this dissemination is to evaluate the effectiveness of MCEQ in measuring the HOTS of pre-service / in-service primary school natural science teachers.

## Instrument

### a. Item Construction

The developed MCEQ was designed based on natural science learning outcomes in primary teacher education. There were two learning outcomes that 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.

### b. Experts' Judgment

In addition to the test, the MCEQ quality was also assessed by experts using the Delphi technique. The experts were asked to assess the aspects of MCEQ 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 MCEQ.

## Data Analysis

The data obtained from the results of the validation test by experts and respondents were analyzed as a basis 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. Qualitative data analysis technique can be carried out by:

- collecting the data in the form of notes, comments, criticisms, and advice from experts which are obtained from the distribution of assessment questionnaires;
- collecting, selecting and classifying data based on test groups; and
- analyzing data and drawing conclusions from various results of the analysis to be used as a basis for taking action to improve the product being developed.

In quantitative data analysis, descriptive and inferential statistics are used. The experts' assessment of the quality of the MCEQ is based on the criteria described in Table 1.

Table 1. MCEQ Quality Criteria Guidelines

Categories	Test Results (scale 100)	Criteria	Action Taken
4	81-100	Very good	Implementation
3	61-80	Good	Implementation
2	41-60	Fair	Revision
1	< 41	Poor	Revision

**Commented [A5]:** There is a lot of information in this section. You don't need to explain the formulation of the analysis. Instead, just write down the techniques you've applied for validity and reliability with a short paragraph.

If the score is  $\geq 60$  (good / very good), an empirical test to obtain construct validity will be carried out. The results from the test are analyzed to determine validity, reliability, discrimination index, difficulty index, and distractor efficiency.

### Validity

The validity of multiple-choice questions is obtained from the formula of point-biserial correlation  $r_{bis} = \frac{M_p - M_q}{S_t} \sqrt{pq}$ . The formula consists of  $r$  = point-biserial correlation coefficient,  $M_p$  = number of respondents who answered correctly,  $M_q$  = number of respondents who answered incorrectly,  $S_t$  = standard deviation for all items,  $P$  = proportion of respondents who answered the question correctly, and  $Q$  = proportion of respondents who answered the question incorrectly. On the other hand, the validity of essay questions is obtained from product-moment correlation, formulated as

$$r_{xy} = \frac{N \sum x_1 y_1 - (\sum x_1)(\sum y_1)}{\sqrt{[N \sum x_1^2 - (\sum x_1)^2][N \sum y_1^2 - (\sum y_1)^2]}}$$

It is indicated that  $r_{xy}$ =correlation between  $x$  dan  $y$ ,  $x_1$ = value of the first  $x$ ,  $y_1$  =value of the first  $y$ , and  $N$  = number of value. A question is considered valid if the value of  $r$  that is calculated ( $r$  count) is greater ( $>$ ) than the value of  $r$  in the statistic table.

### Reliability

The reliability of multiple-choice questions is obtained from the KR-20 formula  $r_{KR-20} = \frac{k}{k-1} \left( \frac{1 - \sum pq}{s^2} \right)$ . It is explained that  $r_{KR-20}$ = correlation coefficient with KR20;  $k$  = number of question items;  $p$ = proportion of correct answer on a particular item;  $q$ = proportion of incorrect answer on a particular item; and  $s^2$ = variance of the total score. On the other hand, the reliability of essay questions is obtained from the product-moment formula  $r_{11} = \left( \frac{n}{n-1} \right) \left( \frac{1 - \sum si^2}{\sum st^2} \right)$ . It is described that  $r_{11}$ = reliability coefficient of the test;  $n$ = number of question items;  $si^2$ = item variance; dan  $st^2$ = total variance. The criteria for reliability are as the following: 0.91–1.00 (very high); 0.71– 0.90 (high); 0.41– 0.70 (moderate); 0.21– 0.40 (low); and Negative – 0.20 (very low).

### ***Discrimination Index***

The discrimination index of multiple-choice questions is obtained from the formula  $DI = \frac{2(KA-KB)}{n}$ . It is described that DI = discrimination index; KA = number of students in the upper group who got the item correct; KB = number of students in the lower group who got the item correct; dan n = number of students. On the other hand, the discrimination index of essay questions is obtained from  $DI = \frac{Mean A - Mean B}{Skor\ maximum}$ . It is explained that DI= discrimination index; Mean A = mean of upper group students; Mean B = mean of lower group students; dan Skor maximum = maximum score of each item. The criteria for discrimination index are as the following: 0.71–1.00 (very different); 0.41–0.70 (different); 0.21–0.40 (fairly different); dan 0.00–0.20 (less different).

### ***Difficulty Index***

The difficulty index of multiple-choice questions is obtained from the formula  $DIF = \frac{JB}{n}$ . It is explained that DIF = difficulty index; JB= number of students who got the item correct; n = number of students. On the other hand, the difficulty index of essay questions is obtained from the formula  $DIF = \frac{Mean}{Maximum\ score}$ . It is described that DIF = difficulty index; Mean= mean of the score; Maximum score = maximum score of each item. The criteria for difficulty index are as the following: 0.71–1.00 (easy); 0.31–0.70 (moderate); and 0.00–0.30 (difficult).

### ***Distractor Efficiency***

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 distractor functions if it is chosen by at least 5% of the testee (Hingorjo & Jaleel, 2012).

## **FINDINGS**

This research has succeeded in developing three MCEQ sets to measure the HOTS of pre-service primary school natural science teachers through the stages of define, design, develop, and dissemination. At the define stage, the urgency of developing MCEQ is based on the high need for HOTS measurement instruments for students. 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, MCEQ is a solution to solve this problem. Furthermore, an analysis of learning outcomes is carried out and the material of the human respiratory system ~~were-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 2.

*Table 2. Analysis of Learning Outcomes and Indicators*

<b>Materials</b>	<b>Course Learning Outcomes</b>	<b>Indicators</b>
<b>Organ Systems</b>	Students are able to 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

The next step after the define stage is the develop stage. At this stage, the blueprint for question items which is presented in Table 3 was designed.

*Table 3. Examples of Blueprint for Question items to Measure HOTS*

<b>Learning Outcomes</b>	<b>Learning Indicators</b>	<b>Question Item Indicators</b>	<b>Number of Question Items</b>	<b>Stimulus</b>	<b>HOTS Level</b>
Students are able to understand the structure and functions of the organs of the respiratory system	Analysing 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
<b>Etc...</b>					

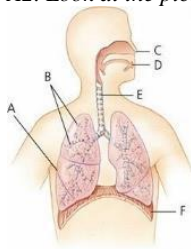
The guidelines above were formulated in the following questions.

*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 pleural membrane
- e. have two lobes

A2. Look at the picture below!



A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...

Source: [artikelmateri.com](http://artikelmateri.com)

- 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 lungs. Explain the reasons!

Answer: .....

Each question has a different stimulus in the form of the 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 clear answers from the students. For multiple-choice questions, each correct answer is given a score of 1, while the score for essay question is 6.

### Validity

The development stage was conducted by developing the guidelines into question items, testing content validity, and conducting an empirical test on the product. The content validity test involved experts in natural science education, learning evaluation, and language. 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 4.

Table 4. Results of Product Assessment by Experts

No	Validators	Value	Qualifications
1	Evaluation experts	79 %	Good
2	Language experts	83.3 %	Very Good
3	Natural science education experts	81.3 %	Very Good
	<b>Average</b>	81.2 %	Very Good

The content validity shows an average value of 81.2%, meaning 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 test results are described in Table 5.

Table 5. Validity Test Result

Question	Package A		Package B		Package C	
	N: 29, R <sub>table</sub> : 0,367		N: 25, R <sub>table</sub> : 0,367		N: 27, R <sub>table</sub> : 0,367	
	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria
A1	0,550	Valid	0,445	Valid	0,443	Valid
A2	0,498	Valid	0,510	Valid	0,474	Valid
A3	0,487	Valid	0,617	Valid	0,403	Valid
A4	0,511	Valid	0,442	Valid	0,426	Valid
A5	0,476	Valid	0,730	Valid	0,599	Valid
A6	0,431	Valid	0,401	Valid	0,497	Valid
A7	0,387	Valid	0,474	Valid	0,500	Valid
A8	0,387	Valid	0,570	Valid	0,409	Valid
A9	0,397	Valid	0,401	Valid	0,705	Valid
A10	0,479	Valid	0,467	Valid	0,416	Valid



Question	Package A		Package B		Package C	
	N: 29, R <sub>table</sub> : 0,367		N: 25, R <sub>table</sub> : 0,367		N: 27, R <sub>table</sub> : 0,367	
	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria
<b>B1</b>	0,693	Valid	0,785	Valid	0,548	Valid
<b>B2</b>	0,608	Valid	0,517	Valid	0,286	<b>Invalid</b>
<b>B3</b>	0,746	Valid	0,474	Valid	0,743	Valid
<b>B4</b>	0,796	Valid	0,471	Valid	0,203	<b>Invalid</b>
<b>B5</b>	0,900	Valid	0,794	Valid	0,470	Valid

Based on Table 5, all items in question set A and B are valid, but in question set C, two items are invalid. A question is said to be valid if it measures what it is intended to measure. An invalid test produces data that is irrelevant to the measurement objective. This can be caused by the difficulty index of the question, distractor function, use of 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 invalid essay 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 analysing 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	A problem is presented, students can identify the shortness of breath 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 bear it. 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**

The reliability test of MCEQ 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 (Beck, Keddy, & Cohen, 1994). The test is said to be reliable or consistent if the scores are similar when the test is taken several times. This research used two methods for the reliability test.

- a) *Kuder-Richardson 20* is a special form of *Cronbach's alpha*. The value ranges from 0-1, with value closes to 1 indicating reliability. This method is used to find the internal consistency coefficient of multiple-choice questions (Quaigrain & Arhin, 2017).
- b) *Cronbach's alpha* is a method that will be used in analyzing essay questions. It is a coefficient of internal consistency and is widely used in social sciences, business, nursing, and other disciplines. It is the average of all split-half reliability estimates of an instrument and is usually used to estimate the reliability of psychometric tests for a sample of testees (Bajpai & Bajpai, 2014).

The results of the reliability test on the three question sets are presented in Table 7.

*Table 7. Results of Questions Reliability Analysis*

<b>Types of Question</b>	<b>Set A Question</b>		<b>Set B Question</b>		<b>Set C Question</b>	
	<b>R<sub>value</sub></b>	<b>Criteria</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
Multiple Choice	0.57	fair	0.67	moderate	0.65	moderate
Essay	0.89	high	0.70	moderate	0.81	high

If the reliability coefficient is within the range of 0.81 to 1.0, it indicates high reliability, 0.61-0.80 indicates moderate reliability, 0.41-0.60 indicates fair reliability, 0.10-0.40 indicates low reliability, and <0.10 indicates that the question is unreliable (Golafshani, 2003).

### **Discrimination Index (DI)**

~~Discrimination~~-The discrimination index is the ability of a test item to distinguish between highly competent testees and those who are not (Panjaitan, Irawati, Sujana, Hanifah, & Djuanda, 2018). Different methods to analyze the discrimination index of objective questions and essay questions were employed. Figure 1 shows the measurement results of the discrimination index for set A, B, and C questions.

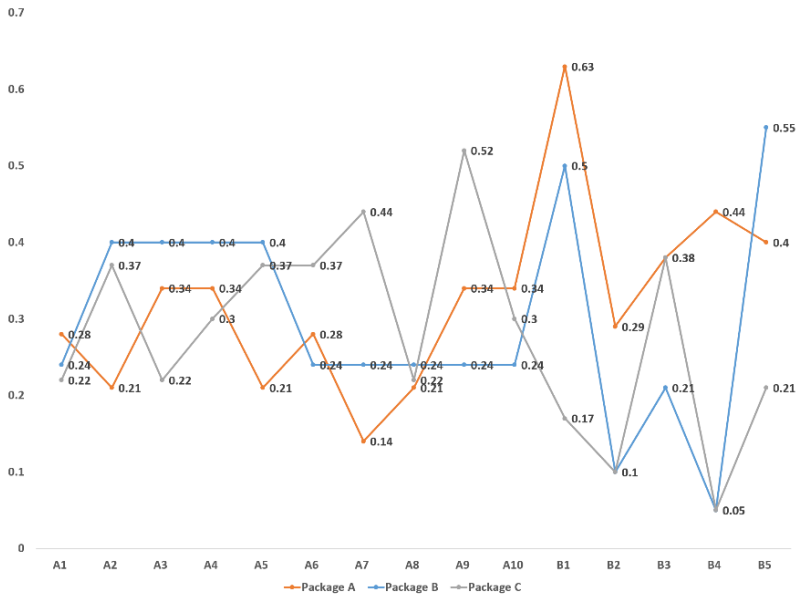


Figure 1. Result of the discrimination index

The data in Figure 1 were categorized based on the ability of the question items to distinguish testee. From these results, the question item is said to have a very good discrimination index if it has a DI of 0.71-1.00; good discrimination index if it has a DI of 0.41-0.70; sufficient discrimination index if it has a DI of 0.21-0.40; poor discrimination index if it has a DI of 0.00-0.20; and if the DI is negative, all question items are said to be bad [26]. The categorization of the Discrimination Index of the question items is shown in Table 8.

Table 8. The categorization of the discrimination index

Package	Discrimination index			
	poor	sufficient	good	very good
A	A7, A9	A1, A3, A4, A5, A6, A8, B1, B2, B3, B4	A2, A10	-
B	A4, A6, A8, B2, B3, B4	A1, A2, A7, A9, B1, B5	A5, A10	A3
C	A2, A10, B1, B2, B3, B4, B5	A4, A8, A9	A1, A3, A5, A6, A7	-

**Difficulty Index (DIF )**

~~Difficulty~~ *The difficulty index* is a measurement of the difficulty index of a question (Karelia, Professor, Pillai, & Vegada, 2013). Analyzing the difficulty index of questions means classifying questions into easy, moderate, and difficult (Chauhan, Chauhan, Chauhan, Vaza, & Rathod, 2015). In constructing test items, it should be noted that a balanced difficulty index should be used. The results of the difficulty index measurement are presented in Figure 2.

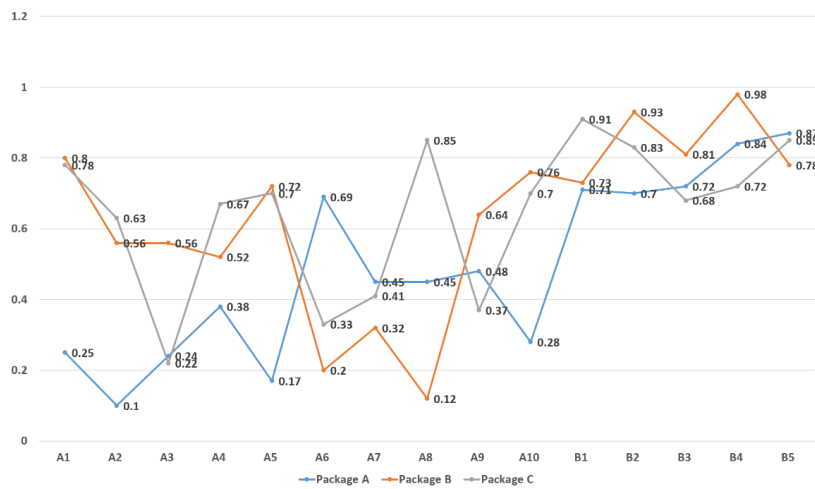


Figure 2. Result of Diffilucty Index (DIF)

A question item is called easy if many testees answer it correctly (DIF: 0.71-1.00). It is called sufficient if there is a balanced number of testees who answer it correctly and incorrectly (DIF: 0.31-0.70). It is called difficult if few testees answer it correctly (DIF: 0.00-0.30). In addition, it is called too easy if the value of P is equivalent to 1.00. An appropriate test item generally has a P-value that ranges from 0.15 to 0.85 (Brown & Hudson, 2002). Based on the criteria and figure 2, the question set A has the following proportion of questions: 20% (difficult), 20% (moderate), and 60% (easy). The proportion in question set B shows that 13.33% of questions are classified as difficult, 20% are moderate, and 66.67% is easy. In question set C, 13.33% of questions are classified as difficult, 20% are moderate, and 66.67% is easy.

### *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 testees (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, F, & DeBoer, 2011).

The more testees 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, the question set A contains 26 distractors that functioned effectively and 14 distractors that did not function effectively. Similarly, question set B shows that 26 distractors functioned effectively, and 14 distractors did not function effectively. Question set C shows that 19 distractors functioned effectively, and 21 distractors 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 9.

*Table 9. Example of Distractor Revision*

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
A6	D	Influenza	Pneumonia	Bring answer choice closer to the answer key
	E	Lung cancer	Polyp	Bring answer choice closer to the answer key
<b>Question set B</b>				
A4	E	Exchange of oxygen in the nasal cavity with CO2 in the tissues	Exchange of oxygen in the bronchial cavity with CO2 in the tissues	Bring answer choice closer to the answer key
<b>Question set C</b>				
A3	B	The process of inspiration in the lungs	The process of managing oxygen in the lungs	Bring answer choice closer to the answer key

MCEQ, which had been declared feasible were used to analyze the HOTS of 79 students taking the Primary School Natural Science Learning Development course in the 5<sup>th</sup> semester of primary teacher education. The results of the analysis are presented in Figure 3.

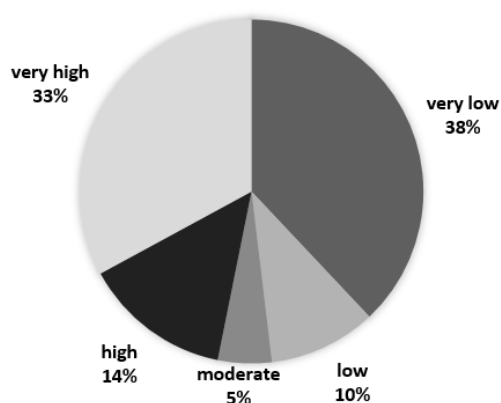


Figure 3. Analysis results of students' HOTS

## DISCUSSION

Evaluation is an activity of identifying, clarifying, and implementing criteria to achieve the success of a program (David, Kartowagiran, & Harjo, 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, Nielsen, & Armstrong, 2013; Widodo & Sudarsono, 2016). Validity indicates what is supposed to be measured by the instrument (Azwar, 2002). 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 MCEQ, natural science education experts and learning evaluation experts were involved.

In this study, the content validity obtained was 81.2% (closer to 1). 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 0.57; 0.65; 0.67; 0.70; 0.81; and 0.89. 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's reliability is higher than 0.70 (Thaneerananon, Triampo, & Nokkaew, 2016). Therefore, the three multiple-choice question sets in this study are considered insufficient to meet the adequacy criteria, while the three 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, Haslam, & Jans, 2013). In order to increase the reliability and validity of items, a number of 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, Estocado, Landers, & Black, 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/environment where the instrument is used (J. O. Chang,



Levy, Seay, & Goble, 2014; Ghosh, Bowles, Ranmuthugala, & Brooks, 2016; Postmes et al., 2013).

Difficulty index (DIF) describes the proportion of students 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 value  $<0.3$  and  $>0.7$  are considered difficult and easy (Khoshaim & Rashid, 2016). DIF has a strong effect on variability in test scores (Thorndike, Cunningham, Thorndike, & Hagen, 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 test, 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 key (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 key. 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, Nagandla, & Agarwal, 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, Ahmad, Aldrees, Khalil, & Ponnampereuma, 2014).

This study provides useful findings that are valuable for the education sector because MCEQ 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 are able to provide an overview of students' thinking skills. The teacher training department can prepare learning activities that can train and empower the prospective teachers' HOTS.

## **CONCLUSION**

This research has succeeded in producing three sets of MCEQ on natural science to measure the higher-order thinking skills of students. Each question set consists of 10 multiple choice questions and five essay questions. Experts' validation shows a very good assessment result. Based on the construct validity test, 43 questions are found to be valid, and two questions are invalid. These invalid questions have been revised based on the item analysis. The reliability test shows that the criteria are sufficient, high, and very high, with Rvalue between 0.57 - 0.89. Most items have an accept difficulty index and a very good discrimination index. The discrimination index is in the moderate to a very high category, while the difficulty index is in the category of very easy, medium, difficult, and very difficult. Revision, particularly on very difficult and very easy questions, was done. The test items that show a very good discrimination index tend to be difficult questions, and items that show poor discrimination index tend to have varied difficulty index. The distractor efficiency shows that 59.2% of distractors are functioning distractors while the remaining 40.8% are non-functioning distractors, which were revised based on the answer analysis of each item. This valid instrument can be developed and implemented by primary teacher education for other courses to identify the HOTS of prospective teachers accurately. The results can reveal the weaknesses of pre-service 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.

## **REFERENCE**

Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *Tes Prestasi: fungsi dan pengembangan pengukuran prestasi belajar*. Yogyakarta: 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>
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- 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. <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.). Boston-USA: 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>
- 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>
- 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. *Evidence-Based Nursing*, 18(3), 66–67. <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.
- Jacobs, L. C., & Chase, C. I. (1992). *Developing and Using Tests Effectively. A Guide for Faculty*. San Francisco, CA: 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. Retrieved from <https://eric.ed.gov/?id=EJ1086950>
- 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., & Samuelson, 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., & Rusydi Rasyid, M. (2016). Studi Kemampuan Mahasiswa Mendesain Perencanaan Pembelajaran Matematika Di Sekolah Menengah Pertama Berbasis Pendekatan Saintifik. *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>

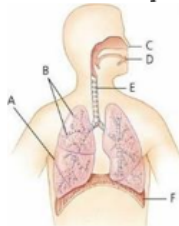
- Panjaitan, R. L., Irawati, R., Sujana, A., Hanifah, N., & Djuanda, D. (2018). Item validity vs. item discrimination index: A redundancy? *Journal of Physics: Conference Series*, 983(1). <https://doi.org/10.1088/1742-6596/983/1/012101>
- Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure 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>
- 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>
- 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>
- Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. Macmillan: 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>

- 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.

## APPENDIXES

### A sample item for the Multiple Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?



# HASIL REVIEW DARI REVIEWER 2

## ROUND 1

### MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primary School Natural Science Teachers

#### ABSTRACT

This study aims to develop MCEQ for measure pre-service primary school natural science teachers' HOTS. This study used a 4-D research design. Evaluation experts, language experts, and natural science education experts were involved in content validation. The quality test conducted by experts showed that the average score of the question was 81.16 (very good). The validity test of the question set A and B demonstrated that all questions were valid. In contrast, in question set C, there were 13 questions classified as valid and two questions classified as invalid. The reliability showed fair, moderate, and high. The discrimination index showed low, moderate, high, and very high. The difficulty index showed very easy, easy, moderate, difficult, and very difficult. The distractor efficiency showed that 59.2% were functioning distractors, and 40.8% were non-functioning. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low.

**Keywords:** MCEQ, instrument, validated, HOTS.

**Commented [WU1]:** Add 1-2 sentences representing the introduction (inform the problem) at the beginning of your abstract

**Commented [WU2]:** The sentence containing the research objective should be in the past tense

**Commented [WU3]:** Add a sentence that contains the implications of your research

## INTRODUCTION

Education in the 21st century requires students to have skills to learn and innovate, to use technology and information media, and to work and survive using life skills. Based on this change in the paradigm of learning in the 21st century, the LPTK (Institute of Teachers' Training) is required to produce qualified prospective teachers. (Bhakti & Maryani, 2017) explained that LPTK has a task to prepare professional teachers, educators for the nation's generation. Teachers are professional occupations that provide expert services and demand academic, pedagogical, social, and professional skills. Teachers are human resources in education who must be able to follow changes quickly (Redhana, 2019). Teachers must be creative, innovative, able to think critically, able to make decisions correctly, and able to solve problems well. These abilities are parts of the teacher's higher-order thinking skills. Therefore, LPTK is expected to be able to produce the best teacher candidates who possess these abilities.

Human resource skills that are demanded in the 21st century are communication, collaboration, critical thinking and problem solving, and creativity and innovation (Arifin, 2017). Students can possess these abilities if the teacher can develop well-planned learning plans. Learning plans that are designed must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rusydi Rasyid, 2016). One approach that meets the purpose is a scientific approach. The implementation of scientific approach in the 2013 curriculum in Indonesia was intended to provide an understanding of getting knowledge of and understanding various materials using scientific approach.

The scientific approach has the potential to maximize HOTS by using scientific reasoning (Pradana, 2020). The scientific approach consists of several main activities, namely observing, questioning, experimenting, associating, communicating, and networking (Pradana, 2020; Susantini, Faizah, Prastiwi, & Suryanti, 2016). All of these scientific activities can potentially influence higher-order thinking (HOTS). HOTS is a thought process that requires 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, making hypotheses to conclude. HOTS is related to cognitive abilities in analyzing, evaluating, and creating.

**Commented [WU4]:** Too many ineffective sentences, poor systematics, and sentences that are too long

**Commented [WU5]:**

**Commented [WU6]:** There are several citation formats that need to be improved

The success of 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 (S. C. Chang & Hwang, 2018; Hartman & Johnson, 2018; He, Holton, Farkas, & Warschauer, 2016; O’Flaherty & Phillips, 2015) and the modification of inquiry with collaboration models have been successfully achieved (Chebii, Wachanga, & Kiboss, 2012; Kovanović, Gašević, Joksimović, Hatala, & Adesope, 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, Rakhmat, & Saepulrohman, 2019), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). In addition, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. (Syafri 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 and is 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) (S. Ahmad et al., 2018). 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. 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. Therefore, this study aims to develop a valid MCEQ in measuring primary teacher education students’ HOTS in natural science. The designed product can be used in many similar institutions to analyze students’ HOTS to be able to find weaknesses and solutions for improvement.

## AIMS

This study aims to develop MCEQ (Multiple Choice and Essay Questions) for measure pre-service primary school natural science teachers’ HOTS.

**Commented [WU7]:** Your analysis related to previous research that examined HOTS in Indonesia is still limited to research published in journals that are not yet internationally reputable (at least have the same quartile as TUSED / Q2). The following is a paper that examines the HOTS of students in Indonesia which is published in Q2 of the journal. Please cite it to strengthen your analysis in Introduction.

*Fauzi, A. and 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>*

In addition, if you are able to obtain previous research that has been published in reputable international journals, I recommend that you cite that paper as well.

**Commented [WU8]:** If the title is pre-service teacher, then the aim is also a pre-service teacher

**Commented [WU9]:** The introduction still does not strongly reflect the importance of your research (why must a pre-service teacher, why must natural science)

## METHODS

### *Participant*

The research subjects consist of subjects for testing and subjects for implementation. In the pilot phase, 81 junior students in primary teacher education were selected to participate. In contrast, in the implementation phase, 75 freshmen who are taking a Natural Science course in the primary teacher education of Universitas Ahmad Dahlan took part in the research. Simple random sampling was used to select participants. The number of samples has met the criteria of sample size in descriptive research.

**Commented [WU10]:** Convey the city and country where the data was collected

### *Development Framework*

This research and development aim to produce a test instrument for 3 question sets in the form of multiple-choice and essay questions as the end product. The final product produced was then tested for measuring the quality through a process of validation and empirical test. In this study, the 4D model by Thiagarajan, Semmel, and Semmel, which includes define, design, develop, and disseminate was employed.

**Commented [WU11]:** Are you using the 4D model as a whole (according to Thiagarajan) or are there any modifications? If there was a modification, tell which stage (or sub-stage) it was modified and why was the modification made? (each sub-stage must appear in the method and be presented in the results)

#### *a. Define*

This define phase is divided into three stages. The first stage is the initial objective analysis. At this stage, the goal of developing a test instrument on natural science material based on higher-level thinking skills was set. The second stage is material analysis. The materials were identified based on the learning outcomes that must be achieved but are considered difficult by students. The third stage is the analysis of the learning outcomes, competence, and learning indicators, which are used to design the question indicators and items. Findings from the define process were used to write the question items in the form of multiple-choice and essay questions (MCEQ.) Both question types were chosen because of their strengths in terms of effectiveness, ease of analysis and practicality in measuring HOTS.

#### *b. Design*

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

- the test questions consist of 3 question sets (A, B, C) and each set has 10 multiple choice and 5 essay questions;
- each HOTS indicator (analysis, evaluation, creation) consists of 3-6 questions that are evenly distributed in each question set;

- one question set contains 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 set, the MCEQ 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 participants.

c. *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 science, experts at evaluation studies, and experts at pedagogical in primary school. The experts were chosen based on their expertise in the related field that corresponds to the product requirements. They were asked to provide suggestions and assess the quality of MCEQ. 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. This experts' assessment was used to repair the instrument. The next process in the develop stage is empirical test. Freshmen and junior students of primary teacher education who are taking a science course became the participants in the test. The test was used to determine validity, reliability, discrimination index, distractor efficiency, and difficulty index. All aspects must meet the high criteria. If any of the parameters receive a low score, it means that improvements will be made in accordance with the results of the item analysis. The final product of the develop phase is a valid MCEQ that meets the experts' judgment and empirical testing. The MCEQ is ready to be implemented in the disseminate stage.

d. *Disseminate*

This phase is the implementation of MCEQ, which has been developed for much wider areas, for example, for students or primary school natural science teachers in other areas. The purpose of this dissemination is to evaluate the effectiveness of MCEQ in measuring the HOTS of pre-service / in-service primary school natural science teachers.

**Commented [WU12]:** Your research aims to develop an instrument for accessing pre-service teacher competencies. However, why at Disseminate do you say that the instrument is also distributed to in-service teachers and students? Is it still reliable even if the subject changes?

**Commented [WU13]:** Your Disseminate Stage lacks the Disseminate described by Thiagarajan

**Commented [WU14]:** Have you also conducted all the activities that you say on Disseminate? I can't find results for that activity in the "Results and Discussion" section

## Instrument

### a. Item Construction

The developed MCEQ was designed based on natural science learning outcomes in primary teacher education. There were two learning outcomes that 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.

### b. Experts' Judgment

In addition to the test, the MCEQ quality was also assessed by experts using the Delphi technique. The experts were asked to assess the aspects of MCEQ 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 MCEQ.

## Data Analysis

The data obtained from the results of the validation test by experts and respondents were analyzed as a basis 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. Qualitative data analysis technique can be carried out by:

- collecting the data in the form of notes, comments, criticisms, and advice from experts which are obtained from the distribution of assessment questionnaires;
- collecting, selecting and classifying data based on test groups; and
- analyzing data and drawing conclusions from various results of the analysis to be used as a basis for taking action to improve the product being developed.

In quantitative data analysis, descriptive and inferential statistics are used. The experts' assessment of the quality of the MCEQ is based on the criteria described in Table 1.

Table 1. MCEQ Quality Criteria Guidelines

Categories	Test Results (scale 100)	Criteria	Action Taken
4	81-100	Very good	Implementation
3	61-80	Good	Implementation
2	41-60	Fair	Revision
1	< 41	Poor	Revision

**Commented [WU15]:** The data analysis you use is too simple. Add a more credible analysis and better explore the quality of instruments, such as the Rasch model or factor analysis

**Commented [WU16]:** State the basics for determining the analysis you use in each analysis you do and also the references you use to categorize the results of the analysis

If the score is  $\geq 60$  (good / very good), an empirical test to obtain construct validity will be carried out. The results from the test are analyzed to determine validity, reliability, discrimination index, difficulty index, and distractor efficiency.

### Validity

The validity of multiple-choice questions is obtained from the formula of point-biserial correlation  $r_{bis} = \frac{M_p - M_q}{S_t} \sqrt{pq}$ . The formula consists of  $r$  = point-biserial correlation coefficient,  $M_p$  = number of respondents who answered correctly,  $M_q$  = number of respondents who answered incorrectly,  $S_t$  = standard deviation for all items,  $P$  = proportion of respondents who answered the question correctly, and  $Q$  = proportion of respondents who answered the question incorrectly. On the other hand, the validity of essay questions is obtained from product-moment correlation, formulated as  $r_{xy} = \frac{N\sum x_1 y_1 - (\sum x_1)(\sum y_1)}{\sqrt{(N\sum x_1^2 - (\sum x_1)^2)(N\sum y_1^2 - (\sum y_1)^2)}}$ . It is indicated that  $r_{xy}$  = correlation between  $x$  dan  $y$ ,  $x_1$  = value of the first  $x$ ,  $y_1$  = value of the first  $y$ , and  $N$  = number of value. A question is considered valid if the value of  $r$  that is calculated ( $r$  count) is greater ( $>$ ) than the value of  $r$  in the statistic table.

**Commented [WU17]:** Is point-biserial correlation able to analyze the validity of items from an instrument? Isn't this analysis more suitable for analyzing discrimination index of items? Can you please inform the references you used?

### Reliability

The reliability of multiple-choice questions is obtained from the KR-20 formula  $r_{KR-20} = \frac{k}{k-1} \left( \frac{1 - \sum pq}{s^2} \right)$ . It is explained that  $r_{KR-20}$  = correlation coefficient with KR20;  $k$  = number of question items;  $p$  = proportion of correct answer on a particular item;  $q$  = proportion of incorrect answer on a particular item; and  $s^2$  = variance of the total score. On the other hand, the reliability of essay questions is obtained from the product-moment formula  $r_{11} = \left( \frac{n}{n-1} \right) \left( \frac{1 - \sum si^2}{\sum st^2} \right)$ . It is described that  $r_{11}$  = reliability coefficient of the test;  $n$  = number of question items;  $si^2$  = item variance; dan  $st^2$  = total variance. The criteria for reliability are as the following: 0.91–1.00 (very high); 0.71– 0.90 (high); 0.41– 0.70 (moderate); 0.21– 0.40 (low); and Negative – 0.20 (very low).

**Commented [WU18]:** State the basics for determining the analysis you use in each analysis you do and also the references you use to categorize the results of the analysis

**Commented [WU19]:** State the basics for determining the analysis you use in each analysis you do and also the references you use to categorize the results of the analysis

### Discrimination Index

The discrimination index of multiple-choice questions is obtained from the formula  $DI = \frac{2(KA-KB)}{n}$ . It is described that DI = discrimination index; KA = number of students in the upper group who got the item correct; KB = number of students in the lower group who got the item correct; dan n = number of students. On the other hand, the discrimination index of essay questions is obtained from  $DI = \frac{Mean A - Mean B}{Skor\ maximum}$ . It is explained that DI = discrimination index; Mean A = mean of upper group students; Mean B = mean of lower group students; dan Skor maximum = maximum score of each item. The criteria for discrimination index are as the following: 0.71–1.00 (very different); 0.41–0.70 (different); 0.21–0.40 (fairly different); dan 0.00–0.20 (less different).

**Commented [WU20]:** State the basics for determining the analysis you use in each analysis you do and also the references you use to categorize the results of the analysis

### Difficulty Index

The difficulty index of multiple-choice questions is obtained from the formula  $DIF = \frac{JB}{n}$ . It is explained that DIF = difficulty index; JB = number of students who got the item correct; n = number of students. On the other hand, the difficulty index of essay questions is obtained from the formula  $DIF = \frac{Mean}{Maximum\ score}$ . It is described that DIF = difficulty index; Mean = mean of the score; Maximum score = maximum score of each item. The criteria for difficulty index are as the following: 0.71–1.00 (easy); 0.31–0.70 (moderate); and 0.00–0.30 (difficult).

**Commented [WU21]:** State the basics for determining the analysis you use in each analysis you do and also the references you use to categorize the results of the analysis

### Distractor Efficiency

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 distractor functions if it is chosen by at least 5% of the testee (Hingorjo & Jaleel, 2012).

### FINDINGS

This research has succeeded in developing three MCEQ sets to measure the HOTS of pre-service primary school natural science teachers through the stages of define, design, develop, and dissemination. At the define stage, the urgency of developing MCEQ is based on the high need for HOTS measurement instruments for students. The instruments that have

**Commented [WU22]:** Present the results of each stage of development (including findings that represent the "Define" and "Dissemination" stages)



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, MCEQ is a solution to solve this problem. Furthermore, an analysis of learning outcomes is carried out and the material of the human respiratory system were 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 2.

*Table 2. Analysis of Learning Outcomes and Indicators*

<b>Materials</b>	<b>Course Learning Outcomes</b>	<b>Indicators</b>
<b>Organ Systems</b>	Students are able to understand the motion system, digestive system, respiratory system, and blood circulatory system	<ul style="list-style-type: none"> <li>. Analyzing the structure and functions of the organs of the respiratory system</li> <li>. Analyzing the respiratory problems experienced by people in the society</li> </ul>

The next step after the define stage is the develop stage. At this stage, the blueprint for question items which is presented in Table 3 was designed.

*Table 3. Examples of Blueprint for Question items to Measure HOTS*

<b>Learning Outcomes</b>	<b>Learning Indicators</b>	<b>Question Item Indicators</b>	<b>Number of Question Items</b>	<b>Stimulus</b>	<b>HOTS Level</b>
Students are able to understand the structure and functions of the organs of the respiratory system	Analysing 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
<b>Etc...</b>					

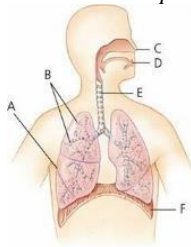
The guidelines above were formulated in the following questions.

*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 pleural membrane
- e. have two lobes

A2. Look at the picture below!



A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...

Source: [artikelmateri.com](http://artikelmateri.com)

- 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 lungs. Explain the reasons!

Answer: .....

Each question has a different stimulus in the form of the 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 clear answers from the students. For multiple-choice questions, each correct answer is given a score of 1, while the score for essay question is 6.

### Validity

The development stage was conducted by developing the guidelines into question items, testing content validity, and conducting an empirical test on the product. The content validity test involved experts in natural science education, learning evaluation, and language. 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 4.

Table 4. Results of Product Assessment by Experts

No	Validators	Value	Qualifications
1	Evaluation experts	79 %	Good
2	Language experts	83.3 %	Very Good
3	Natural science education experts	81.3 %	Very Good
	<b>Average</b>	81.2 %	Very Good

The content validity shows an average value of 81.2%, meaning 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 test results are described in Table 5.

Table 5. Validity Test Result

Question	Package A		Package B		Package C	
	N: 29, R <sub>table</sub> : 0,367		N: 25, R <sub>table</sub> : 0,367		N: 27, R <sub>table</sub> : 0,367	
	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria
A1	0,550	Valid	0,445	Valid	0,443	Valid
A2	0,498	Valid	0,510	Valid	0,474	Valid
A3	0,487	Valid	0,617	Valid	0,403	Valid
A4	0,511	Valid	0,442	Valid	0,426	Valid
A5	0,476	Valid	0,730	Valid	0,599	Valid
A6	0,431	Valid	0,401	Valid	0,497	Valid
A7	0,387	Valid	0,474	Valid	0,500	Valid
A8	0,387	Valid	0,570	Valid	0,409	Valid
A9	0,397	Valid	0,401	Valid	0,705	Valid
A10	0,479	Valid	0,467	Valid	0,416	Valid

Question	Package A		Package B		Package C	
	N: 29, R <sub>table</sub> : 0,367		N: 25, R <sub>table</sub> : 0,367		N: 27, R <sub>table</sub> : 0,367	
	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria	R <sub>value</sub>	Criteria
<b>B1</b>	0,693	Valid	0,785	Valid	0,548	Valid
<b>B2</b>	0,608	Valid	0,517	Valid	0,286	<b>Invalid</b>
<b>B3</b>	0,746	Valid	0,474	Valid	0,743	Valid
<b>B4</b>	0,796	Valid	0,471	Valid	0,203	<b>Invalid</b>
<b>B5</b>	0,900	Valid	0,794	Valid	0,470	Valid

Based on Table 5, all items in question set A and B are valid, but in question set C, two items are invalid. A question is said to be valid if it measures what it is intended to measure. An invalid test produces data that is irrelevant to the measurement objective. This can be caused by the difficulty index of the question, distractor function, use of 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 invalid essay 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 analysing 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	A problem is presented, students can identify the shortness of breath 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 bear it. 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**

The reliability test of MCEQ 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 (Beck, Keddy, & Cohen, 1994). The test is said to be reliable or consistent if the scores are similar when the test is taken several times. This research used two methods for the reliability test.

- a) *Kuder-Richardson 20* is a special form of *Cronbach's alpha*. The value ranges from 0-1, with value closes to 1 indicating reliability. This method is used to find the internal consistency coefficient of multiple-choice questions (Quaigrain & Arhin, 2017).
- b) *Cronbach's alpha* is a method that will be used in analyzing essay questions. It is a coefficient of internal consistency and is widely used in social sciences, business, nursing, and other disciplines. It is the average of all split-half reliability estimates of an instrument and is usually used to estimate the reliability of psychometric tests for a sample of testees (Bajpai & Bajpai, 2014).

The results of the reliability test on the three question sets are presented in Table 7.

*Table 7. Results of Questions Reliability Analysis*

<b>Types of Question</b>	<b>Set A Question</b>		<b>Set B Question</b>		<b>Set C Question</b>	
	<b>R<sub>value</sub></b>	<b>Criteria</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
Multiple Choice	0.57	fair	0.67	moderate	0.65	moderate
Essay	0.89	high	0.70	moderate	0.81	high

If the reliability coefficient is within the range of 0.81 to 1.0, it indicates high reliability, 0.61-0.80 indicates moderate reliability, 0.41-0.60 indicates fair reliability, 0.10-0.40 indicates low reliability, and <0.10 indicates that the question is unreliable (Golafshani, 2003).

### **Discrimination Index (DI)**

Discrimination index is the ability of a test item to distinguish between highly competent testees and those who are not (Panjaitan, Irawati, Sujana, Hanifah, & Djuanda, 2018). Different methods to analyze the discrimination index of objective questions and essay questions were employed. Figure 1 shows the measurement results of the discrimination index for set A, B, and C questions.

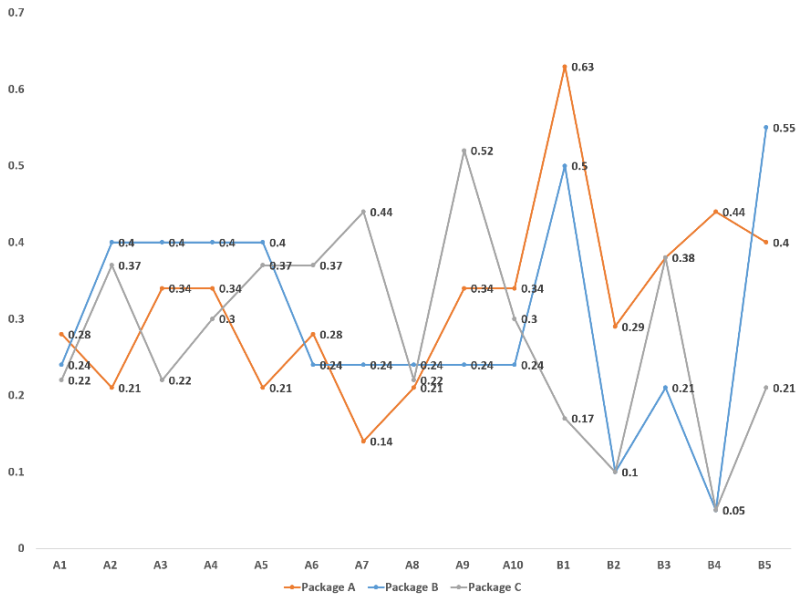


Figure 1. Result of the discrimination index

The data in Figure 1 were categorized based on the ability of the question items to distinguish testee. From these results, the question item is said to have a very good discrimination index if it has a DI of 0.71-1.00; good discrimination index if it has a DI of 0.41-0.70; sufficient discrimination index if it has a DI of 0.21-0.40; poor discrimination index if it has a DI of 0.00-0.20; and if the DI is negative, all question items are said to be bad [26]. The categorization of the Discrimination Index of the question items is shown in Table 8.

Table 8. The categorization of discrimination index

Package	Discrimination index			
	poor	sufficient	good	very good
A	A7, A9	A1, A3, A4, A5, A6, A8, B1, B2, B3, B4	A2, A10	-
B	A4, A6, A8, B2, B3, B4	A1, A2, A7, A9, B1, B5	A5, A10	A3
C	A2, A10, B1, B2, B3, B4, B5	A4, A8, A9	A1, A3, A5, A6, A7	-

**Difficulty Index (DIF )**

Difficulty index is a measurement of the difficulty index of a question (Karelia, Professor, Pillai, & Vegada, 2013). Analyzing the difficulty index of questions means classifying questions into easy, moderate, and difficult (Chauhan, Chauhan, Chauhan, Vaza, & Rathod, 2015). In constructing test items, it should be noted that a balanced difficulty index should be used. The results of the difficulty index measurement are presented in Figure 2.

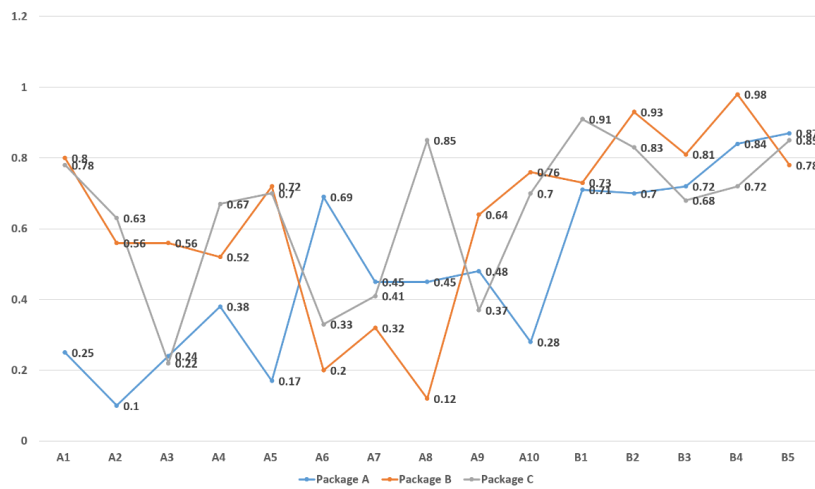


Figure 2. Result of Diffilucty Index (DIF)

A question item is called easy if many testees answer it correctly (DIF: 0.71-1.00). It is called sufficient if there is a balanced number of testees who answer it correctly and incorrectly (DIF: 0.31-0.70). It is called difficult if few testees answer it correctly (DIF: 0.00-0.30). In addition, it is called too easy if the value of P is equivalent to 1.00. An appropriate test item generally has P-value that ranges from 0.15 to 0.85 (Brown & Hudson, 2002). Based on the criteria and figure 2, the question set A has the following proportion of questions: 20% (difficult), 20% (moderate), and 60% (easy). The proportion in question set B shows that 13.33% of questions are classified as difficult, 20% are moderate, and 66.67% is easy. In question set C, 13.33% of questions are classified as difficult, 20% are moderate, and 66.67% is easy.



### ***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 testees (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, F, & DeBoer, 2011).

The more testees 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, the question set A contains 26 distractors that functioned effectively and 14 distractors that did not function effectively. Similarly, question set B shows that 26 distractors functioned effectively, and 14 distractors did not function effectively. Question set C shows that 19 distractors functioned effectively, and 21 distractors 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 9.

*Table 9. Example of Distractor Revision*

<b>Number of Question Item</b>	<b>Answer Choices</b>	<b>Distractors</b>		<b>Purpose of Revision</b>
		<b>Before Revision</b>	<b>After Revision</b>	
<b>Question set A</b>				
<b>A6</b>	D	Influenza	Pneumonia	Bring answer choice closer to the answer key
	E	Lung cancer	Polyp	Bring answer choice closer to the answer key
<b>Question set B</b>				
<b>A4</b>	E	Exchange of oxygen in the nasal cavity with CO2 in the tissues	Exchange of oxygen in the bronchial cavity with CO2 in the tissues	Bring answer choice closer to the answer key
<b>Question set C</b>				
<b>A3</b>	B	The process of inspiration in the lungs	The process of managing oxygen in the lungs	Bring answer choice closer to the answer key

MCEQ, which had been declared feasible were used to analyze the HOTS of 79 students taking the Primary School Natural Science Learning Development course in the 5<sup>th</sup> semester of primary teacher education. The results of the analysis are presented in Figure 3.

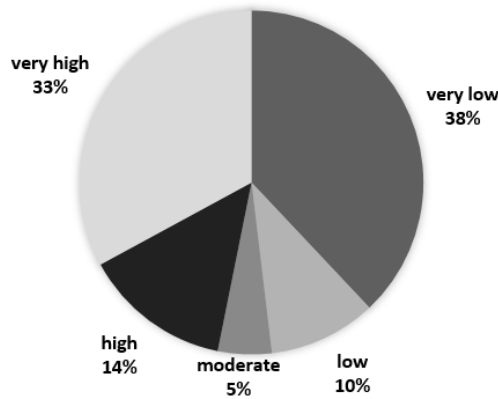


Figure 3. Analysis results of students' HOTS

## DISCUSSION

Evaluation is an activity of identifying, clarifying, and implementing criteria to achieve the success of a program (David, Kartowagiran, & Harjo, 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, Nielsen, & Armstrong, 2013; Widodo & Sudarsono, 2016). Validity indicates what is supposed to be measured by the instrument (Azwar, 2002). 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 MCEQ, natural science education experts and learning evaluation experts were involved.

In this study, the content validity obtained was 81.2% (closer to 1). This means that the validity index agreement is higher than the items in the instrument, which are appropriate

**Commented [WU23]:** Discuss the urgency of measuring pre-service HOTS using the instruments you have developed

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 0.57; 0.65; 0.67; 0.70; 0.81; and 0.89. 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's reliability is higher than 0.70 (Thaneerananon, Triampo, & Nokkaew, 2016). Therefore, the three multiple-choice question sets in this study are considered insufficient to meet the adequacy criteria, while the three 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, Haslam, & Jans, 2013). In order to increase the reliability and validity of items, a number of 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, Estocado, Landers, & Black, 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/environment where the instrument is used (J. O. Chang,

Levy, Seay, & Goble, 2014; Ghosh, Bowles, Ranmuthugala, & Brooks, 2016; Postmes et al., 2013).

Difficulty index (DIF) describes the proportion of students 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 value  $<0.3$  and  $>0.7$  are considered difficult and easy (Khoshaim & Rashid, 2016). DIF has a strong effect on variability in test scores (Thorndike, Cunningham, Thorndike, & Hagen, 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 test, 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 key (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 key. 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, Nagandla, & Agarwal, 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, Ahmad, Aldrees, Khalil, & Ponnampereuma, 2014).

This study provides useful findings that are valuable for the education sector because MCEQ 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 are able to provide an overview of students' thinking skills. The teacher training department can prepare learning activities that can train and empower the prospective teachers' HOTS.

**Commented [WU24]:** Convey the limitation of your study

**Commented [WU25]:** Also inform future studies that you plan or recommend

## CONCLUSION

This research has succeeded in producing three sets of MCEQ on natural science to measure the higher-order thinking skills of students. Each question set consists of 10 multiple choice questions and five essay questions. Experts' validation shows a very good assessment result. Based on the construct validity test, 43 questions are found to be valid, and two questions are invalid. These invalid questions have been revised based on the item analysis. The reliability test shows that the criteria are sufficient, high, and very high, with Rvalue between 0.57 - 0.89. Most items have an accept difficulty index and a very good discrimination index. The discrimination index is in the moderate to very high category, while the difficulty index is in the category of very easy, medium, difficult, and very difficult. Revision, particularly on very difficult and very easy questions, was done. The test items that show a very good discrimination index tend to be difficult questions, and items that show poor discrimination index tend to have varied difficulty index. The distractor efficiency shows that 59.2% of distractors are functioning distractors while the remaining 40.8% are non-functioning distractors, which were revised based on the answer analysis of each item. This valid instrument can be developed and implemented by primary teacher education for other courses to identify the HOTS of prospective teachers accurately. The results can reveal the weaknesses of pre-service teachers' HOTS so that the institution can develop learning models that lead to the empowerment of HOTS.

**Commented [WU26]:** Divide into two paragraphs. The first paragraph concludes your research according to the objectives stated earlier. The second paragraph contains recommendations and implications based on your research process and findings

## 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.

## REFERENCE

Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, G. (2014). The relationship between non-functioning distractors and item difficulty of multiple-choice

**Commented [WU27]:** 1. Some of your reference metadata need to be improved  
Add references from reputable international journals, one of which I recommend in the introduction

- 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. (2002). *Tes Prestasi: fungsi dan pengembangan pengukuran prestasi belajar*. Yogyakarta: 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>
- Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- 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. <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.). Boston-USA: 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>
- 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>
- 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. *Evidence-Based Nursing*, 18(3), 66–67. <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.
- Jacobs, L. C., & Chase, C. I. (1992). *Developing and Using Tests Effectively. A Guide for Faculty*. San Francisco, CA: 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. Retrieved from <https://eric.ed.gov/?id=EJ1086950>
- 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., & Samuelson, 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., & Rusydi Rasyid, M. (2016). Studi Kemampuan Mahasiswa Mendesain Perencanaan Pembelajaran Matematika Di Sekolah Menengah Pertama Berbasis Pendekatan Saintifik. *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>

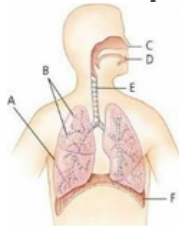
- Panjaitan, R. L., Irawati, R., Sujana, A., Hanifah, N., & Djuanda, D. (2018). Item validity vs. item discrimination index: A redundancy? *Journal of Physics: Conference Series*, 983(1). <https://doi.org/10.1088/1742-6596/983/1/012101>
- Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure 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>
- 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>
- 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>
- Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. Macmillan: 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>

- 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.

## APPENDIXES

### A sample item for the Multiple Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# REVISI ROUND 1 KE-1

## HMCEQ (HOTs Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Teacher Student

Ika Maryani<sup>1</sup>, Zuhdan Kun Prasetyo<sup>2</sup>, Insih Wilujeng<sup>3</sup>, Siwi Purwanti<sup>4</sup>, Meita Fitrianawati<sup>5</sup>

<sup>1</sup> Assist. Prof., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id), ORCID ID: [0000-0002-7154-2902](https://orcid.org/0000-0002-7154-2902)

<sup>2</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [zuhdan@uny.ac.id](mailto:zuhdan@uny.ac.id), ORCID ID: [0000-0001-9342-1565](https://orcid.org/0000-0001-9342-1565)

<sup>3</sup> Assoc. Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [insih@uny.ac.id](mailto:insih@uny.ac.id), ORCID ID: [0000-0003-1900-7985](https://orcid.org/0000-0003-1900-7985)

<sup>4</sup> Assist. Prof., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [siwi.purwanti@pgsd.uad.ac.id](mailto:siwi.purwanti@pgsd.uad.ac.id), ORCID ID: [0000-0002-1433-7531](https://orcid.org/0000-0002-1433-7531)

<sup>5</sup> Assist. Prof., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [meita.fitrianawati@pgsd.uad.ac.id](mailto:meita.fitrianawati@pgsd.uad.ac.id), ORCID ID: [0000-0002-3748-3718](https://orcid.org/0000-0002-3748-3718)

### Correspondent

Ika Maryani, M.Pd. Universitas Ahmad Dahlan, Jl. Ki Ageng Pemanahan 19 Sorosutan, Yogyakarta-Indonesia, +6282297575204, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id).

### ABSTRACT

HOTS is a very crucial thinking skill needed by prospective teachers to develop 21st century learning. This study aimed to develop HMCEQ to measure the higher order thinking skills of teacher students of elementary school education department. This study used a 4-D model by Thiagarajan. Experts at natural science, experts at evaluation studies, and experts at pedagogical in primary school were involved in content validation. There were 156 teacher students involved as test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ questions, each of which consisted of 10 multiple choice and 5 essays. The validity test by Rasch Model showed that there were 7 multiple choice questions classified as fit, and 3 questions (number 3, 9, and 10) were classified as misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. Reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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% were functioning distractors, and 40.8% were non-functioning. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze the teacher student HOTS. This data can be used as the basis for developing a program to increase the competence of prospective teachers.

**Keywords:** HMCEQ, instrument, higher order thinking skills.

**Commented [A1]: Reviewer A:** The abbreviation of the name of your instrument cannot fully explain what it is measuring. Multiple choice and essay questions for what? Higher order thinking skills?

**Commented [A2R1]:** I added HOTS in the abbreviation of the instrument name

**Commented [A3]: Reviewer A:** The abstract is insufficient to cover the overall study. It is not clear how many questions were prepared for the instrument and to whom.

**Commented [A4]:** reviewer B: Add 1-2 sentences representing the introduction (inform the problem) at the beginning of your abstract

**Commented [A5]:** reviewer B: The sentence containing the research objective should be in the past tense

**Commented [A6]:** reviewer B: Add a sentence that contains the implications of your research

**Commented [A7]: Reviewer A:** please write open

## INTRODUCTION

The 21<sup>st</sup> century education requires students to have life skill, innovative, creative, adaptive, and technology literate. Based on this change, an Institute of Teachers' Training is required to produce qualified prospective teachers. Bhakti & Maryani (2017) explained that they have an important task to prepare professional teachers. Teachers are professional occupations that provide expert service and demand academic, pedagogical, social and professional skills. They must be able to follow changes quickly (Redhana, 2019) and also need to be creative, innovative, able to think critically, be able to make correct decisions, and be able to solve problems well. These abilities are parts of the teacher's higher-order thinking skills. Therefore, teacher training is expected to be able to produce the best teacher candidates 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. The learning plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). One approach that meets the purpose is a scientific approach. 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 maximize 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 higher-order thinking skills (HOTS). HOTS is a thought process that requires 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, making hypotheses to conclude. It is related to cognitive abilities in analyzing, evaluating, and creating.

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 (S. C. 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 &

**Commented [A8]:** reviewer B: Too many ineffective sentences, poor systematics, and sentences that are too long

**Commented [A9]:** reviewer B: There are several citation formats that need to be improved

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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). In addition, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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 and is 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) (S. Ahmad et al., 2018).

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. Therefore, this study aims to develop a valid HMCEQ in measuring the higher order thinking skills of teacher students of 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.

## AIMS

This study aims to develop HMCEQ (HOTS Multiple Choice and Essay Questions) to measure the higher order thinking skills of teacher students of elementary school education department.

## METHODS

### *Participant*

**Commented [A10]:** reviewer B: Your analysis related to previous research that examined HOTS in Indonesia is still limited to research published in journals that are not yet internationally reputable (at least have the same quartile as TUSED / Q2). The following is a paper that examines the HOTS of students in Indonesia which is published in Q2 of the journal. Please cite it to strengthen your analysis in Introduction.

*Fauzi, A. and 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>*

In addition, if you are able to obtain previous research that has been published in reputable international journals, I recommend that you cite that paper as well.

**Commented [A11]:** Reviewer A: The introduction still does not strongly reflect the importance of your research (why must a pre-service teacher, why must natural science)

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 junior students in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 freshmen 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. The number of samples has met the criteria of sample size in descriptive research.

**Commented [A12]:** Reviewer A: Convey the city and country where the data was collected

### ***Development Framework***

This research and development 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 was employed.

**Commented [A13]:** Reviewer A: Are you using the 4D model as a whole (according to Thiagarajan) or are there any modifications? If there was a modification, tell which stage (or sub-stage) it was modified and why was the modification made? (each sub-stage must appear in the method and be presented in the results)

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) consists of 1-3 questions;
- the instruments contains 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 teacher students.

*c. 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 develop stage is the empirical test. 156 Freshmen and junior students of primary teacher education who are taking a science course 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 develop phase is a valid HMCEQ that meets the experts' judgment and empirical testing. The HMCEQ is ready to be implemented in the disseminate stage.

*d. 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 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.

## **Instrument**

*a. Item Construction*

The developed HMCEQ was designed based on natural science learning outcomes in primary teacher education. There were two learning outcomes that 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.

*b. 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 basis 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 empirical test to determine validity, reliability, discrimination index, and difficulty index. The quality of instrument are analyzed by Rasch Model. 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 distractor functions if it is chosen by at least 5% of the testee (Hingorjo & Jaleel, 2012).

**Commented [A14]: Reviewer A:** There is a lot of information in this section. You don't need to explain the formulation of the analysis. Instead, just write down the techniques you've applied for validity and reliability with a short paragraph.

**Commented [A15]: reviewer B:**The data analysis you use is too simple. Add a more credible analysis and better explore the quality of instruments, such as the Rasch model or factor analysis

### **FINDINGS**

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

#### **Define**

At the define stage, the urgency of developing HMCEQ is based on the high need for HOTS measurement instruments for students. 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 2.

**Commented [A16]: reviewer B:** Present the results of each stage of development (including findings that represent the "Define" and "Dissemination" stages)

Table 1. Analysis of Learning Outcomes and Indicators

Materials	Course Learning Outcomes	Indicators
Organ Systems	Students are able to understand the motion system, digestive system, respiratory system, and blood circulatory system	<ul style="list-style-type: none"> <li>. Analyzing the structure and functions of the organs of the respiratory system</li> <li>. Analyzing the respiratory problems experienced by people in the society</li> </ul>

### Design

Tahap design menghasilkan buku instrument yang berisi kisi-kisi, kumpulan soal yang terdiri dari 10 pilihan ganda dan 5 esay), petunjuk pengerjaan soal, lembar jawab, kunci jawaban, dan panduan penskoran. At this stage, the blueprint for question items which is presented in Table 3 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	HOTS Level
Students are able to understand the structure and functions of the organs of the respiratory system	Analysing 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
<b>Etc...</b>					

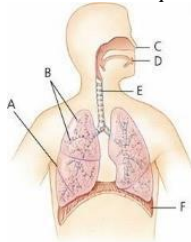
The guidelines above were formulated in the following questions.

#### 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 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 is 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 lungs. Explain the reasons!*

*Answer: .....*

Each question has a different stimulus in the form of the 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 clear answers from the students. For multiple-choice questions, each correct answer is given a score of 1, while the score for essay question is 6.

**Develop**

**Validity**

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 4.

*Table 3. Results of Product Assessment by Experts*

No	Validators	Value	Qualifications
1	Evaluation experts	79 %	Good
2	Pedagogical in primary schoolexperts	83.3 %	Very Good
3	Natural science experts	81.3 %	Very Good
<b>Average</b>		81.2 %	Very Good

The content validity shows an average value of 81.2%, meaning 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	Aitem	X <sup>2</sup>	Pr (> X <sup>2</sup> )	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	fit
	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 <sub>value</sub>	Criteria
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}}(0,367) = \text{valid}$*

Based on Table 4 and 5, 3 items in multiple choices question are misfit and 7 items are fit, whereas 2 items in essay question 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 analysing 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	A problem is presented, students can identify the shortness of breath 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 bear it. 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***

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). In this study, the cronbach alpha coefficient of multiplechoice questions is 0.605 (reliable) and the essay questions is 0.61 (reliable).

### ***Discrimination Index (DI) and Difficulty Index (DIF )***

The discrimination index is the ability of a test item to distinguish between highly competent testees 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 discriminat items is as follows.  $D \geq 0.4$  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  $D \leq 0$ , 20 questions were discarded / bad questions (Vishnumolakala et al., 2016). The results of the difficulty index multiple choice questions showed in Figure 2 and the essay ones in Figure 3.

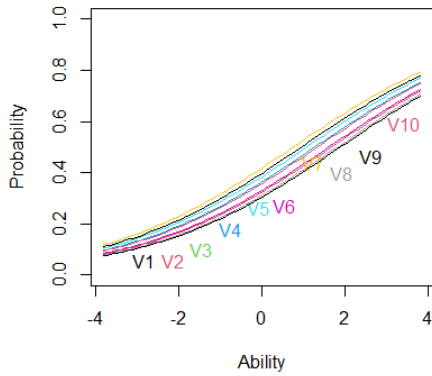


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions

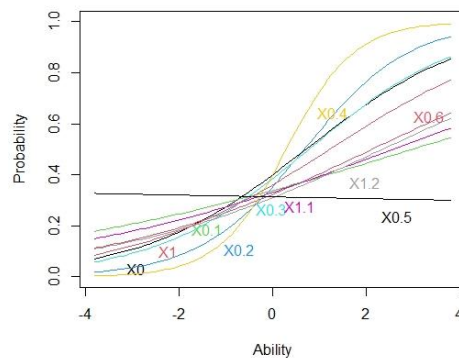


Figure 2. Result of Difficulty Index (DIF) of essay questions

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. Difficulties index and discriminant index data are shown in Table 7.

Table 7. difficulties index and discriminant index of questions

Type of questions	Number	Difficulties index	Kesimpulan	Discriminant Index	Kesimpulan
Multiple Choice Questions	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 Questions	1.	-3,542	Esay	0.219	Sufficient
	2.	-2,631	Esay	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 testees (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 testees 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, there are 26 distractors that 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 9.

Table 8. Example of Distractor Revision

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
Question set A 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 teacher students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

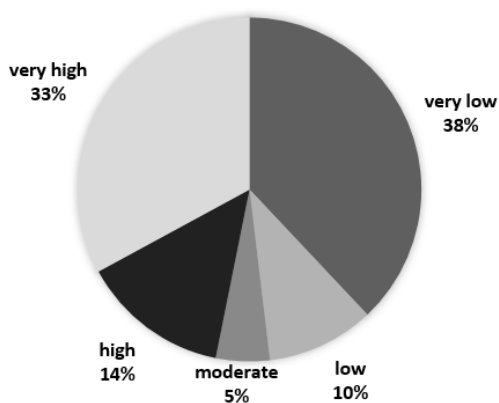


Figure 3. Analysis results of students' HOTS

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

### Disseminate

The instrument has been complete 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, 2002). 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

**Commented [A17]:** reviewer B: Discuss the urgency of measuring pre-service HOTS using the instruments you have developed

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's is reliable (Thaneerananon et al., 2016). Therefore, the multiple-choice question in this study are 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). In order to increase the reliability and validity of items, a number of 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/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 students 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 value  $<0.3$  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 test, 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 key (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 key. 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 are able to provide an overview of students' 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 science material used in the instrument is limited to respiratory system. Therefore, it is necessary to develop instruments in other materials.

## CONCLUSION

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teacher students. The instrument consists of 10 multiple choice

**Commented [A18]:** reviewer B: Convey the limitation of your study

**Commented [A19]:** reviewer B: Also inform future studies that you plan or recommend

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 misfit. 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 index. The distractor efficiency shows that 59.2% of distractors are functioning distractors while the remaining 40.8% are non-functioning distractors, 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 teacher students' 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.

#### REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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).

**Commented [A20]:** reviewer B: Divide into two paragraphs. The first paragraph concludes your research according to the objectives stated earlier. The second paragraph contains recommendations and implications based on your research process and findings

**Commented [A21]:** reviewer B: Some of your reference metadata need to be improved. Add references from reputable international journals, one of which I recommend in the introduction

- Azwar, S. (2002). *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>
- 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.
- 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., & Samuelson, 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>
- 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>
- 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>

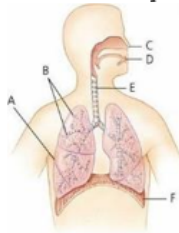
Young, D. L., Estocado, N., Landers, M. R., & Black, J. (2011). A pilot study providing evidence for the validity of a new tool to improve assignment of national pressure ulcer advisory panel stage to pressure ulcers. *Advances in Skin & Wound Care*, 24(4), 168–175. <https://doi.org/10.1097/01.asw.0000396304.90710.ea>

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.

## APPENDIXES

### A sample item for the Multiple Choice Question (C5)

A2. Look at the picture below!



A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...

Source: [artikelmateri.com](http://artikelmateri.com)

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# REVISI ROUND 1 KE-2

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of **Teachers Training Students**

---

### ABSTRACT

HOTS is a very crucial thinking skill needed by prospective teachers to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the students 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 teachers training student as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple-choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze teachers training student' HOTS. **This data can be used as the reference for developing competency improvement programs for teachers training students, for example through HOTS-oriented learning models and HOTS improvement training for teachers training student. The teacher training department can prepare learning activities that can train and empower their students' HOTS.**

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

## INTRODUCTION

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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, HOTS is 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. 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 elementary school education department in science learning. What is meant by valid here is that it has been through testing by experts and empirical tests. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of elementary school education department in science learning. This instrument can be used to see the students' HOTS, so that teachers training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## **AIMS**

This study aims to develop a valid HMCEQ 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**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 **students in their 2<sup>nd</sup> year** in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 **students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) **refers to Bloom taxonomy**, consists of 1-3 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 teachers training students.

#### *c. 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 primary teacher education department who are taking a natural science course to become 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.

#### *d. 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.

## **Instrument**

### *a. 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.

#### *b. 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 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 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, develop, and dissemination.

### **Define**

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*

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

## **Design**

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*

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

*Etc...*

---

The guidelines above were formulated in the following questions.

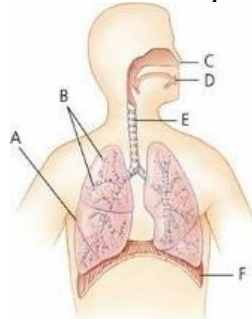
*Multiple Choice Questions*

*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 the 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 question 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

## Develop

### Validity

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

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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*

<b>Item</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}} (0,367) = \text{valid}$*

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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

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). **From the Rasch analysis, the Cronbach's alpha of multiple-choice questions is 0.605 (reliable) and the essay questions are 0.61 (reliable). This reliability value is sufficient and may be used for further research.**

## Discrimination Index (DI) and Difficulty Index (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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

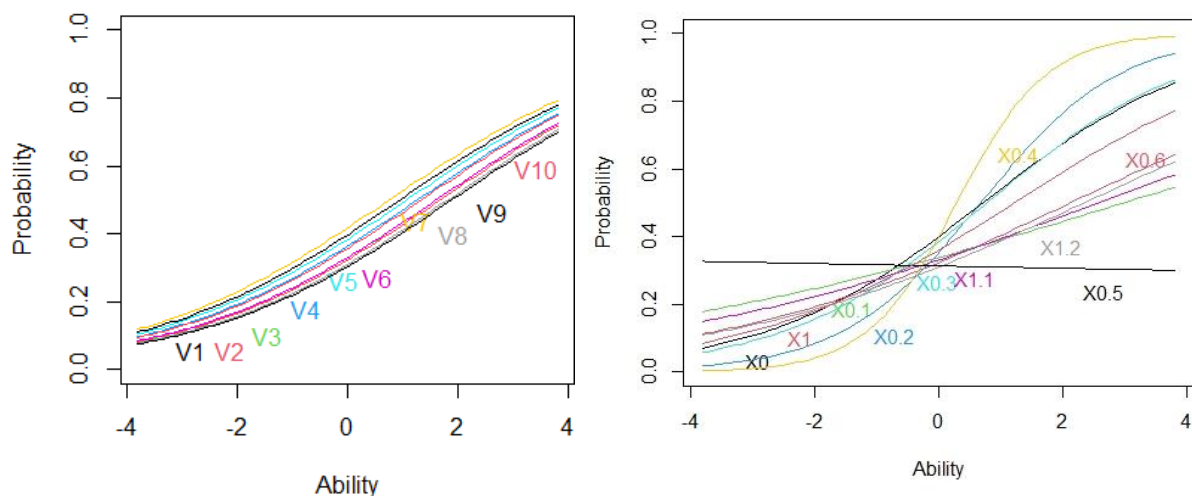


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions      Figure 2. Result of Difficulty Index (DIF) of essay questions

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
<b>Multiple Choice</b>	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
<b>Essay</b>	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. Example of Distractor Revision

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
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 teachers training students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

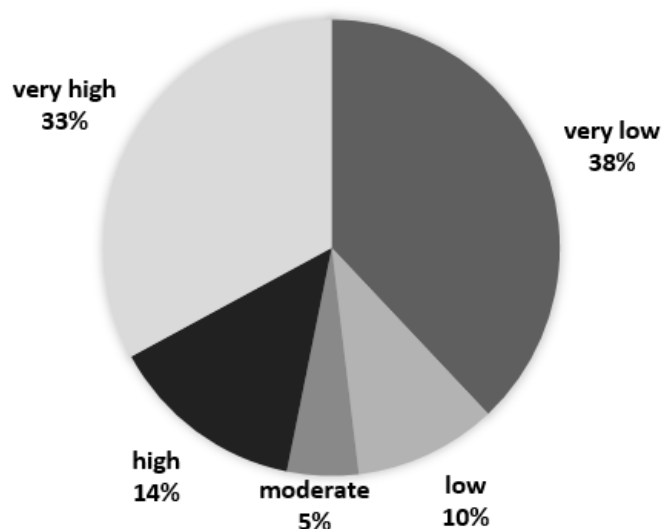


Figure 3. Analysis results of teachers training students' HOTS

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

### Disseminate

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, 2002). The validity of the instrument can be identified from content validity and empirical test on each question item (Lissitz & Samuelson, 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 teachers training students 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 key (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 key. 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  $1\text{NFD} < 2\text{NFD} < 3\text{NFD}$ . 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 teachers training students' 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

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of **teachers training students**. 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 misfit. 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 index. 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 teachers training students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *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>
- 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.
- 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., & Samuelson, 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>
- Sugiyanta, S., & Soenarto, S. (2016). An evaluation model of educational quality assurance at

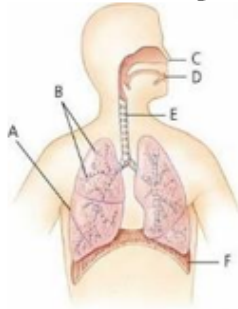


- junior high schools. *Research and Evaluation in Education*, 2(2), 194. <https://doi.org/10.21831/reid.v2i2.11118>
- Susantini, E., Faizah, U., Prastiwi, M. S., & Suryanti. (2016). Developing educational video to improve the use of scientific approach in cooperative learning. *Journal of Baltic Science 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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# REVISI ROUND 1 KE-3

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of **Teachers Training Students**

---

### ABSTRACT

HOTS is a very crucial thinking skill needed by prospective teachers to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the students 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 teachers training student as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple-choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze teachers training student' HOTS. **This data can be used as the reference for developing competency improvement programs for teachers training students, for example through HOTS-oriented learning models and HOTS improvement training for teachers training student. The teacher training department can prepare learning activities that can train and empower their students' HOTS.**

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

## INTRODUCTION

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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, HOTS is 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. 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 elementary school education department in science learning. What is meant by valid here is that it has been through testing by experts and empirical tests. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of elementary school education department in science learning. This instrument can be used to see the students' HOTS, so that teachers training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## **AIMS**

This study aims to develop a valid HMCEQ 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**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 **students in their 2<sup>nd</sup> year** in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 **students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) **refers to Bloom taxonomy**, consists of 1-3 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 teachers training students.

#### *c. 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 primary teacher education department who are taking a natural science course to become 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.

#### *d. 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.

## **Instrument**

### *a. 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.

#### *b. 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 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 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, develop, and dissemination.

### **Define**

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*

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

## **Design**

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*

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

*Etc...*

---

The guidelines above were formulated in the following questions.

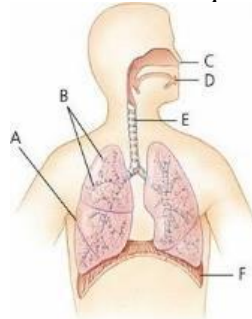
*Multiple Choice Questions*

*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 the 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 question 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

## Develop

### Validity

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

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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*

<b>Item</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}} (0,367) = \text{valid}$*

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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**

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). **From the Rasch analysis, the Cronbach's alpha of multiple-choice questions is 0.605 (reliable) and the essay questions are 0.61 (reliable). This reliability value is sufficient and may be used for further research.**

### **Discrimination Index (DI) and Difficulty Index (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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

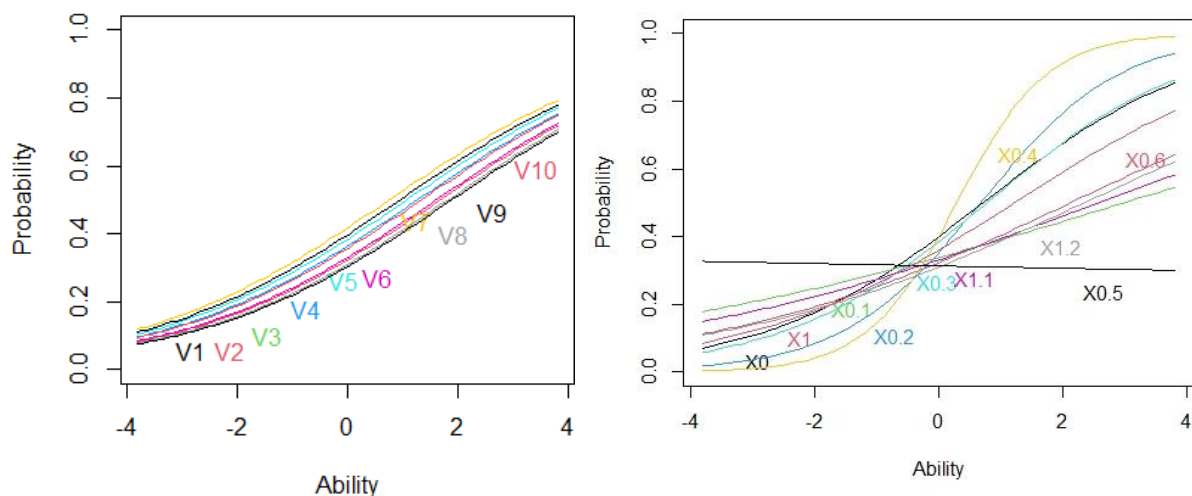


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions      Figure 2. Result of Difficulty Index (DIF) of essay questions

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
<b>Multiple Choice</b>	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
<b>Essay</b>	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. Example of Distractor Revision

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
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 teachers training students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

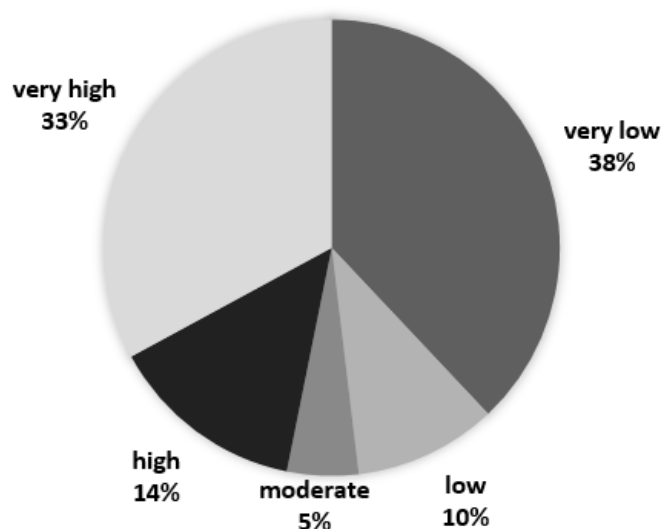


Figure 3. Analysis results of teachers training students' HOTS

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

### Disseminate

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, 2002). The validity of the instrument can be identified from content validity and empirical test on each question item (Lissitz & Samuelson, 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 teachers training students 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 key (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 key. 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  $1\text{NFD} < 2\text{NFD} < 3\text{NFD}$ . 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 teachers training students' 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

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of **teachers training students**. 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 misfit. 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 index. 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 teachers training students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *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>
- 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.
- 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., & Samuelson, 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>
- Sugiyanta, S., & Soenarto, S. (2016). An evaluation model of educational quality assurance at

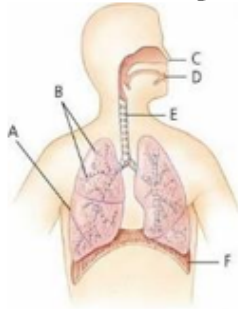
- junior high schools. *Research and Evaluation in Education*, 2(2), 194. <https://doi.org/10.21831/reid.v2i2.11118>
- Susantini, E., Faizah, U., Prastiwi, M. S., & Suryanti. (2016). Developing educational video to improve the use of scientific approach in cooperative learning. *Journal of Baltic Science 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.



## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: [artikelmateri.com](http://artikelmateri.com)*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?


[← Back to Submissions](#)

1018 / Ika Maryani et al. / HOTS Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills

Library

Workflow

Publication

# BUKTI PROSES REVIEW ROUND 2

Submission

Review

Copyediting

Production

Round 1

Round 2

## Round 2 Status

Submission accepted.

## Notifications

<a href="#">[tused] Editor Decision</a>	2020-09-26 11:46 AM
<a href="#">[tused] Editor Decision</a>	2020-09-26 11:48 AM
<a href="#">[tused] Editor Decision</a>	2021-02-17 08:47 AM
<a href="#">[tused] Editor Decision</a>	2021-04-06 09:29 AM
<a href="#">[tused] Editor Decision</a>	2021-12-31 11:23 AM

## Reviewer's Attachments

2757 Reviewer #1	January 13, 2021
2982 Reviewer #2	February 12, 2021

## Revisions



▶  3320 Article Text, rev 20 maret 2021.docx	March 20, 2021	Article Text
▶  3367 Other, author response form.docx	March 20, 2021	Other
▶  3470 Article Text, article revision-march 29.docx	March 29, 2021	Article Text
▶  3498 Article Text, article revision-March 31.docx	March 31, 2021	Article Text
▶  3501 Accept Submission-Article Text, article revision-April 1.docx	April 1, 2021	Article Text
▶  4191 Article Text, article revision-TUSED-Ika Maryani-proofread.docx	June 10, 2021	Article Text

## Review Discussions

Name	From	Last Reply	Replies	Closed
<a href="#">Status of your manuscript</a>	ideveci 2020-12-01 08:47 AM	ikamaryani 2020-12-15 02:15 AM	2	<input type="checkbox"/>
▶ <a href="#">information</a>	ikamaryani 2021-01-20 07:04 AM	-	0	<input type="checkbox"/>
▶ <a href="#">article revision</a>	ikamaryani 2021-03-16 07:00 AM	ideveci 2021-03-27 10:54 AM	3	<input type="checkbox"/>
▶ <a href="#">article revision</a>	ikamaryani 2021-03-29 07:17 AM	ikamaryani 2021-04-01 09:39 AM	4	<input type="checkbox"/>

99+ **Compose**

Mail **Inbox** 4,580

Chat **Starred**

Spaces **Snoozed**

**Important**

**Sent**

Meet **Drafts** 132

**Categories**

**More**

**Labels**

UAD

Q tused

UNIVERSITAS  
AHMAD DAHLAN

27 of 43

Dear Ika Maryani Maryani,

We have reached a decision regarding your submission to Journal of Turkish Science Education " **MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primary School Natural Science Teachers**". Your article was evaluated by two reviewers in Round 2. We have reached the "**Revision Required**" decision regarding your manuscript. In addition, both reviewers expressed their opinions on the fulltext (in attached files, if there is). Thus, the reviewers and editorial views pointed out some serious shortcomings and inadequacies in your work. If you think you can make these corrections, please revise your manuscript.

Our decision is to: "**Revision Required**"

**Important note:**

1. Please indicate **in color** on the fulltext corrections you have made for the opinions of both the referees and the editor (if there is).
2. Also, upload your detailed answers to the reviewers' and editorial opinions (if there is) as "**author response form**" (as a separate word file in detail) to the system.

Once again, thank you for submitting your manuscript to Journal of Turkish Science Education and we look forward to receiving your revision **in four weeks (Until 17 April 2021)**.

**Reviewer #1** (You can see the details on the attached full text)

-This article focuses on '**A Validated Instrument to Measure Higher Order Thinking Skills of Student-Teacher**'

-which is crucial to check and assess student teachers' knowledge on HOTS. However, the authors have to be consistent in addressing the respondents of the study

-as it is very confusing (student teachers OR prospective teachers or freshmen).

**Reviewer #2** (You can see the details on the attached full text)

**HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Student-Teacher**

**Abstract:**

1. Student-Teacher??? is this a prospective teacher or student and teacher???
2. use consistent terms for the subject prospective teacher, student-teacher, or student?
3. **Introduction :**
4. use consistent terms for the subject prospective teacher, student-teacher, or student?
5. Please clarify the study of "HOTS" in this research using bloom taxonomy or Marzano?
6. **Method**
7. the meaning junior students in participant
8. Describe the sampling technique
9. Rasch Model does this apply to multiple choices or both? This instrument consisted of multiple choice

# HASIL REVIEW DARI REVIEWER 1

## ROUND 2

### HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Student-Teacher

---

#### ABSTRACT

HOTS is a very crucial thinking skill needed by prospective teachers to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of student 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 student teachers as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple-choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze student teachers' HOTS. This data can be used as the basis for developing a program to increase the competence of prospective teachers.

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

**Commented [A1]:** Why program? What does the program refer to?  
Please explain. Why not for prospective teachers' HOTS.

## INTRODUCTION

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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. Therefore, teacher training is expected to be able to produce the best teacher candidates 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 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 (S. C. 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-

**Commented [A2]:** My suggestion is to have one section to discuss on Higher-Order Thinking Skills established by Bloom's Taxonomy and its relation in education and to your study. Reformation in HOTS as we have revised taxonomy as well.

grade social science multiple choice questions (Yuniar et al., 2019), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. Ahmad et al., 2018).

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. Therefore, this study aims to develop a valid HMCEQ in measuring the higher-order thinking skills of student teachers 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.

## AIMS

This study aims to develop HMCEQ (HOTS Multiple Choice and Essay Questions) to measure the higher-order thinking skills of student teachers of the elementary school education department.

## METHODS

### *Participant*

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 junior students in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 freshmen who are taking a Natural Science course in the primary teacher education of Universitas Ahmad Dahlan, Yogyakarta, Indonesia took part in

**Commented [A3]:** Please be consistent using the term student teachers or prospective teachers. In your abstract, you mentioned prospective teachers.

**Commented [A4]:** Therefore, this study aims to develop a valid HMCEQ in measuring the higher-order thinking skills of student teachers 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. OR This study aims to develop HMCEQ (HOTS Multiple Choice and Essay Questions) to measure the higher-order thinking skills of student teachers of the elementary school education department.

Please be precise.

**Commented [A5]:** Freshmen, do they refer to student teachers/prospective teachers, please be consistent

the research. Simple random sampling was used to select participants. The number of samples has met the criteria of sample size in descriptive research.

**Commented [A6]:** I suggest you support your reasons for selecting random sampling from scholars like Creswell, etc to show rigour in your methodology

### ***Development Framework***

This research and development 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 was employed.

#### *a. 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.

#### *b. 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 HOTS indicator (analysis, evaluation, creation) consists of 1-3 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 teacher students.

#### *c. 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 freshmen and junior students of primary teacher education who are taking a science course to become 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.

*d. 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.

## **Instrument**

*a. 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.

*b. 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 basis 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 was analyzed by the Rasch Model. 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 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 higher-order thinking skills of teacher students of elementary school education department through the stages of define, design, develop, and dissemination.

#### **Define**

At the defined stage, the urgency of developing HMCEQ is based on the high need for HOTS measurement instruments for students. 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.

**Commented [A7]:** Student teachers right?

Table 1. Analysis of Learning Outcomes and Indicators

<b>Materials</b>	<b>Course Learning Outcomes</b>	<b>Indicators</b>
Organ Systems	Students can understand the motion system, digestive system, respiratory system, and blood circulatory system	<ul style="list-style-type: none"> <li>Analyzing the structure and functions of the organs of the respiratory system</li> <li>Analyzing the respiratory problems experienced by people in the society</li> </ul>

**Design**

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

<b>Learning Outcomes</b>	<b>Learning Indicators</b>	<b>Question Item Indicators</b>	<b>Number of Question Items</b>	<b>Stimulus</b>	<b>HOTS Level</b>
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
<b>Etc...</b>					

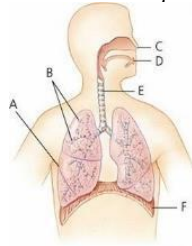
The guidelines above were formulated in the following questions.

*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 the 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 question is 6.

**Develop**

**Validity**

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

No	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
<b>Average</b>		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, distractor 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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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	fit
	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 <sub>value</sub>	Criteria
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

Note:  $R > R_{table} (0,367) = \text{valid}$

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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

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). In this study, the Cronbach Alpha coefficient of multiple-choice questions is 0.605 (reliable) and the essay questions are 0.61 (reliable).

### Discrimination Index (DI) and Difficulty Index (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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

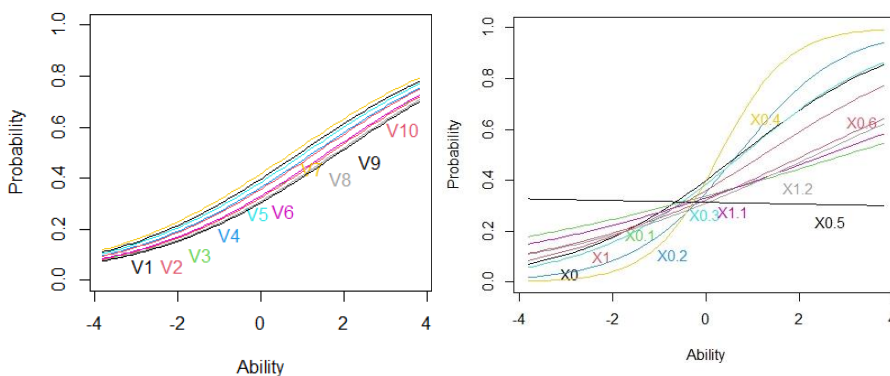




Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions  
 Figure 2. Result of Difficulty Index (DIF) of essay questions

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	Esay	0.219	Sufficient
	2.	-2,631	Esay	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. Example of Distractor Revision

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
Question set A 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 teacher students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

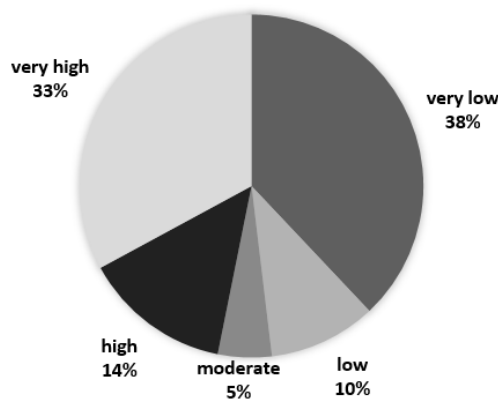


Figure 3. Analysis results of students' HOTS

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

### Disseminate

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, 2002). 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 students 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 key (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 key. 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 students' 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

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teacher students. 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 misfit. 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 index. 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.

**Commented [A8]:** You can take this example to replace the program you mentioned in your abstract. This explanation is more clear.

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 teacher students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *Tes Prestasi: fungsi dan pengembangan pengukuran prestasi belajar*. Pustaka Pelajar.
- Bajpai, R., & Bajpai, S. (2014). The 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.000000000000016>
- 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 a 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>
- 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 the 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.
- 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 the 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., & Samuelson, K. (2007). Further Clarification Regarding Validity and Education. *Educational Researcher*, 36(8), 482–484.



<https://doi.org/10.3102/0013189x07311612>

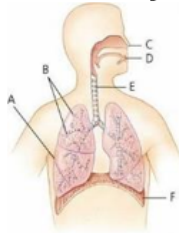
- Mayordomo, R. M., & Onrubia, J. (2015). Work coordination and collaborative knowledge construction in a small group collaborative virtual task. *Internet and Higher Education*, 25, 96–104. <https://doi.org/10.1016/j.iheduc.2015.02.003>
- Nur Aini, D. F., & Sulistyani, N. (2019). Pengembangan Instrumen 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>
- 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>
- 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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

A2. Look at the picture below!



A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...

Source: [artikelmateri.com](http://artikelmateri.com)

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# HASIL REVIEW DARI REVIEWER 2 ROUND 2

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Student-Teacher

**Commented [A1]:** Student-Teacher??? is this a prospective teacher or student and teacher???

### ABSTRACT

HOTS is a very crucial thinking skill needed by prospective teachers to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of student 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 student teachers as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple-choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze student teachers' HOTS. This data can be used as the basis for developing a program to increase the competence of prospective teachers.

**Commented [A2]:** consistent use of terms prospective teachers or student-teacher

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

## INTRODUCTION

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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. Therefore, teacher training is expected to be able to produce the best teacher candidates who possess these abilities.

Commented [A3]: using same term : prospective teachers

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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. Ahmad et al., 2018).

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. Therefore, this study aims to develop a valid HMCEQ in measuring the higher-order thinking skills of student teachers 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.

## AIMS

This study aims to develop HMCEQ (HOTS Multiple Choice and Essay Questions) to measure the higher-order thinking skills of student teachers of the elementary school education department.

## METHODS

### *Participant*

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 junior students in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 freshmen who are taking a Natural Science course in the primary teacher education of Universitas Ahmad Dahlan, Yogyakarta, Indonesia took part in

**Commented [A4]:** the background is dominated by learning that measures HOTS, there is no apparent urgency on the importance of developing the instruments developed

**Commented [A5]:** same size font

**Commented [A6]:** the meaning junior students????

**Commented [A7]:** describe the sampling technique,

the research. Simple random sampling was used to select participants. The number of samples has met the criteria of sample size in descriptive research.

**Commented [A8]:** the subject is students or prospective teachers?

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) consists of 1-3 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 teacher students.

#### ***c. 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 freshmen and junior students of primary teacher education who are taking a science course to become 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.

*d. 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.

## **Instrument**

*a. 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.

*b. 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 basis 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 was analyzed by the Rasch Model. 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 distractor functions if it is chosen by at least 5% of the testing participants (Hingorjo & Jaleel, 2012).

**Commented [A9]:** does this apply to multiple choices or both?  
This instrument consisted of multiple choice questions and essays  
Describe in more detail how to analyze the data

**Commented [A10]:** using CTT or IRT ?

### FINDINGS

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

#### Define

At the defined stage, the urgency of developing HMCEQ is based on the high need for HOTS measurement instruments for students. 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

<b>Materials</b>	<b>Course Learning Outcomes</b>	<b>Indicators</b>
Organ Systems	Students can understand the motion system, digestive system, respiratory system, and blood circulatory system	<ul style="list-style-type: none"> <li>Analyzing the structure and functions of the organs of the respiratory system</li> <li>Analyzing the respiratory problems experienced by people in the society</li> </ul>

**Design**

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

<b>Learning Outcomes</b>	<b>Learning Indicators</b>	<b>Question Item Indicators</b>	<b>Number of Question Items</b>	<b>Stimulus</b>	<b>HOTS Level</b>
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
<b>Etc...</b>					

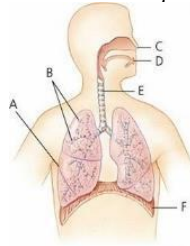
The guidelines above were formulated in the following questions.

*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 the 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 question is 6.

**Commented [A11]:** describe the scoring rubric for essay question ? Politomous?

**Develop**

**Validity**

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

No	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
<b>Average</b>		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, distractor 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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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	fit
	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

Commented [A12]: check with the fit criteria used

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

Table 5. Validity Test Result Essay Questions

Item	R <sub>value</sub>	Criteria
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

Note:  $R > R_{table} (0,367) = \text{valid}$

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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

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). In this study, the Cronbach Alpha coefficient of multiple-choice questions is 0.605 (reliable) and the essay questions are 0.61 (reliable).

**Commented [A13]:** In the rasch model, reliability can be checked with the information function

### Discrimination Index (DI) and Difficulty Index (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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

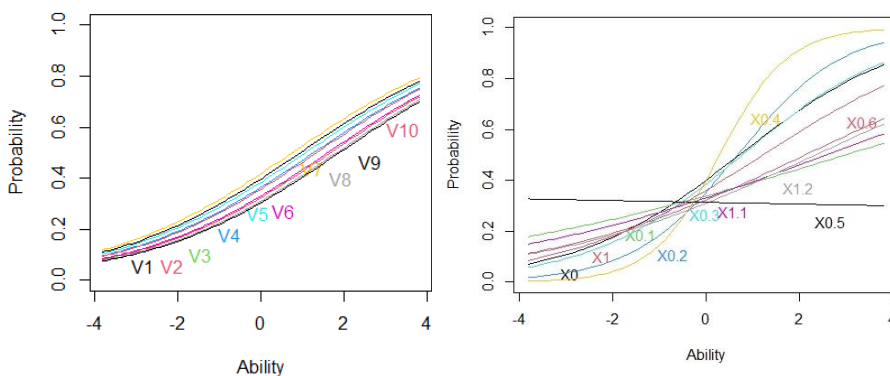


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions      Figure 2. Result of Difficulty Index (DIF) of essay questions

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	Essay	0.219	Sufficient
	2.	-2,631	Essay	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

Commented [A14]: ????

### 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. Example of Distractor Revision

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
Question set A 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 teacher students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

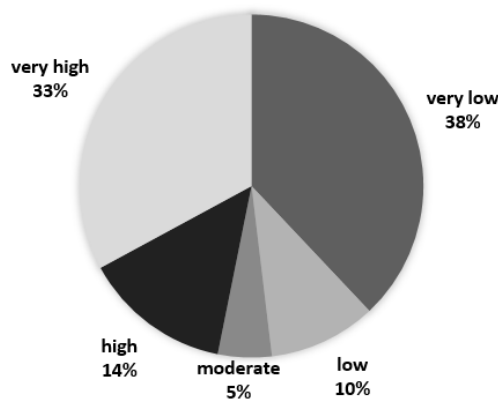


Figure 3. Analysis results of students' HOTS

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

### Disseminate

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, 2002). 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 students 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 key (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 key. 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 students' 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**

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teacher students. 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 misfit. 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 index. 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 teacher students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *Tes Prestasi: fungsi dan pengembangan pengukuran prestasi belajar*. Pustaka Pelajar.
- Bajpai, R., & Bajpai, S. (2014). The 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.000000000000016>
- 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 a 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>
- 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 the 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.
- 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 the 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., & Samuelson, 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>
- Sugiyanta, S., & Soenarto, S. (2016). An evaluation model of educational quality assurance at

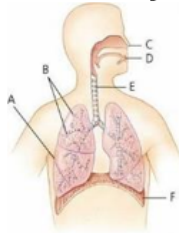


- junior high schools. *Research and Evaluation in Education*, 2(2), 194. <https://doi.org/10.21831/reid.v2i2.11118>
- Susantini, E., Faizah, U., Prastiwi, M. S., & Suryanti. (2016). Developing educational video to improve the use of scientific approach in cooperative learning. *Journal of Baltic Science 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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

A2. Look at the picture below!



A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...

Source: [artikelmateri.com](http://artikelmateri.com)

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# REVISI ROUND 2 KE-1

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Teachers Training Students

---

### ABSTRACT

HOTS is a very crucial thinking skill needed by teachers training students to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the teachers training students 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 teachers training student as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple-choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze teachers training student' HOTS. This data can be used as the reference for developing competency improvement programs for teachers training students, for example through HOTS-oriented learning models and HOTS improvement training for teachers training student. The teacher training department can prepare learning activities that can train and empower their students' HOTS.

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

## **INTRODUCTION**

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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, HOTS is 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. 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 elementary school education department in science learning. What is meant by valid here is that it has been through testing by experts and empirical tests. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of elementary school education department in science learning. This instrument can be used to see the students' HOTS, so that teachers training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## **AIMS**

This study aims to develop a valid HMCEQ 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**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 students in their 2<sup>nd</sup> year in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) refers to Bloom taxonomy, consists of 1-3 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 teachers training students.

#### *c. 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 primary teacher education department who are taking a natural science course to become 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.

#### *d. 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.

## **Instrument**

### *a. 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.

#### *b. 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 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 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, develop, and dissemination.

### **Define**

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*

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

## **Design**

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*

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

*Etc...*

---

The guidelines above were formulated in the following questions.

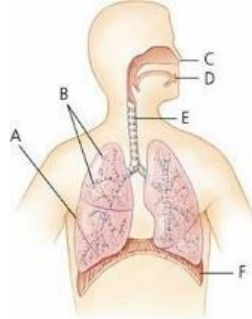
*Multiple Choice Questions*

*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 the 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 question 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

## Develop

### Validity

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*

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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*

<b>Item</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}} (0,367) = \text{valid}$*

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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

The reliability value in the Rasch model is indicated by the value of person separation and item separation. The greater the price of person separation, the better the tests are arranged because the items in it are able to reach individuals with high to low level abilities. item separation indicates how much of the sample subjected to measurement is spread across a linear interval scale. The higher the item separation value, the better the measurement will be. This index is useful for defining the meaning of the constructs we measure (Sumintono & Widhiarso, 2015). 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 indicators that should be observed in the reliability values are: Cronbach alpha ( $\alpha$ ) value (KR-20), person reliability value, person measure, and valid responses (Mohamad et al., 2015)(Shahirah Saidi & Moi Siew, 2019). From the Rasch analysis, the Cronbach's alpha (KR-20) of multiple-choice questions is 0.605 (reliable) and the essay questions are 0.61 (reliable). This reliability value is sufficient and may be used for further research (Sumintono & Widhiarso, 2015).

```
-----  
          X^2 Pr(>X^2)  
V1  22.9292  0.0003  : fit  
V2  12.6841  0.0265  : fit  
V3   5.9195  0.3141  : misfit  
V4  22.6654  0.0004  : fit  
V5  22.5403  0.0004  : fit  
V6   9.2658  0.0989  : fit  
V7  28.5175 <0.0001  : fit  
V8  16.6519  0.0052  : fit  
V9   4.0696  0.5394  : misfit  
V10  8.6818  0.1224  : misfit  
-----  
Cronbach's alpha (KR-20) for the 'X' data-set  
  
Items: 10  
Sample units: 154  
alpha: 0.605  
  
Bootstrap 95% CI based on 154 samples  
2.5% 97.5%  
0.100 0.483  
  
CRONBACH ALPHA (KR-20) 0.605  
-----
```

Figure 1. Result of Reliability Test of Multiple Choices Questions

```

-----
Reliability analysis
Call: alpha(x = data)

raw_alpha std.alpha G6(smc) average_r S/N ase mean sd median_r
0.61 0.62 0.61 0.25 1.6 0.049 3.3 0.93 0.28

lower alpha upper 95% confidence boundaries
0.52 0.61 0.71
-----
Reliability if an item is dropped:
raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
X1 0.64 0.64 0.59 0.31 1.8 0.048 0.0089 0.32
X2 0.51 0.50 0.46 0.20 1.0 0.063 0.0204 0.22
X3 0.58 0.59 0.56 0.26 1.4 0.056 0.0208 0.26
X4 0.50 0.50 0.47 0.20 1.0 0.066 0.0183 0.22
X5 0.56 0.57 0.55 0.25 1.4 0.058 0.0291 0.28
-----
Item statistics
n raw.r std.r r.cor r.drop mean sd
X1 154 0.48 0.51 0.30 0.20 3.4 1.4
X2 154 0.68 0.72 0.64 0.48 3.3 1.3
X3 154 0.61 0.60 0.43 0.34 3.2 1.5
X4 154 0.71 0.71 0.63 0.49 3.4 1.4
X5 154 0.66 0.62 0.45 0.37 3.0 1.7

Non missing response frequency for each item
1 2 3 4 5 6 miss
X1 0.09 0.18 0.31 0.19 0.14 0.08 0
X2 0.07 0.22 0.25 0.27 0.14 0.05 0
X3 0.18 0.14 0.25 0.21 0.13 0.09 0
X4 0.08 0.23 0.26 0.14 0.21 0.06 0
X5 0.28 0.20 0.09 0.16 0.21 0.06 0
-----

```

Figure 2. Result of Reliability Test of Essay Questions

### ***Discrimination Index (DI) and Difficulty Index (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  $-1.00 \text{ logit} < b < +1.00 \text{ logit}$ , 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.  $D \geq 0.4$  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  $D \leq 0, 20$  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.

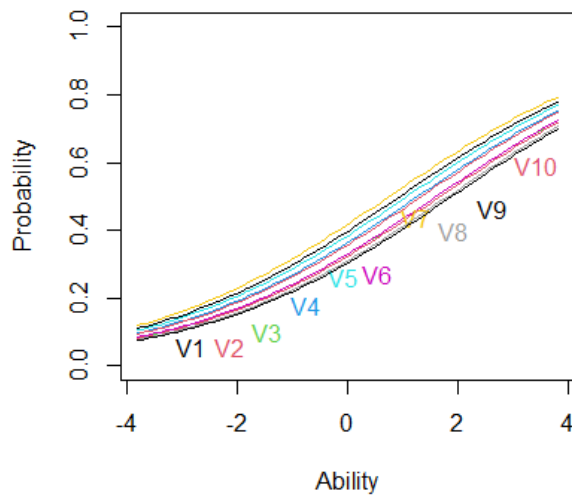


Figure 3. Result of Difficulty Index (DIF) of multiple-choice questions

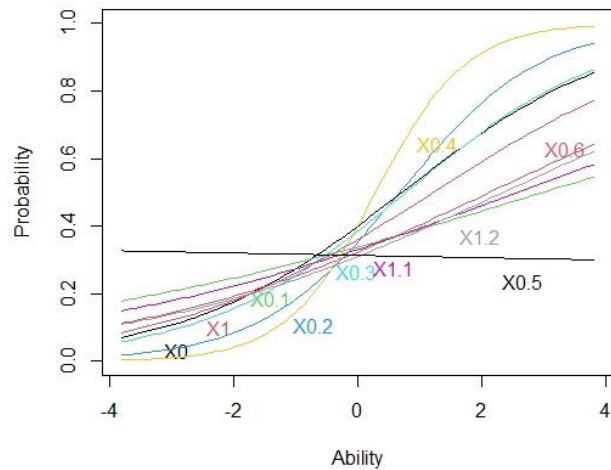


Figure 4. Result of Difficulty Index (DIF) of essay questions

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 &

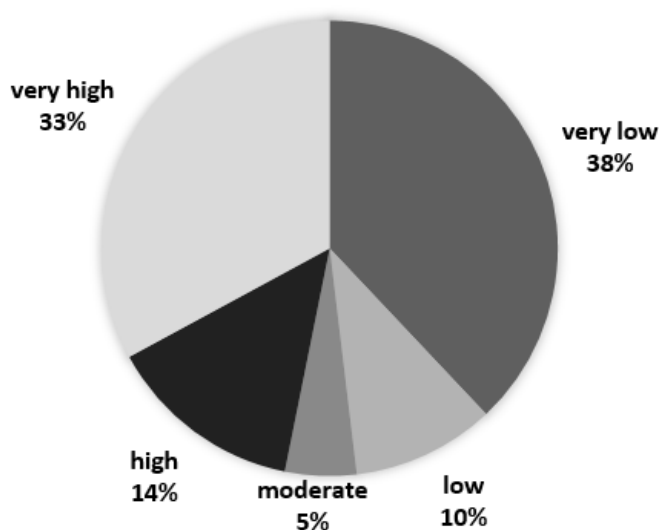


Jaleel, 2012). The effectiveness of the distractor is how well the wrong option can deceive the testees who do not know the correct answer (Herrmann-Abell et al., 2011). The more testing participants were choosing the distractor, the more it functions appropriately. A good distractor will be chosen evenly 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. Example of Distractor Revision*

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
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 teachers training students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.



*Figure 5. Analysis results of teachers training students' HOTS*

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

### **Disseminate**

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, 2002). 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 teachers training students 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 key (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 key. 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 teachers training students' 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**

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teachers training students. 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 misfit. 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 index. 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 teachers training students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *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>
- 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.
- 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., & Samuelson, 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>
- Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the Validity and Reliability of Research Instruments. *Procedia - Social and Behavioral Sciences*, 204, 164–171. <https://doi.org/10.1016/j.sbspro.2015.08.129>
- 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>
- Shahirah Saidi, S., & Moi Siew, N. (2019). Reliability and Validity Analysis of Statistical Reasoning Test Survey Instrument using the Rasch Measurement Model. *International Electronic Journal of Mathematics Education*, 14(3), 535–546.  
<https://doi.org/10.29333/iejme/5755>
- 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–

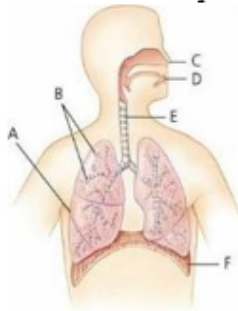
319. <https://doi.org/10.1039/c5rp00214a>

- 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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: [artikelmateri.com](http://artikelmateri.com)*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# REVISI ROUND 2 KE-2

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Teachers Training Students

---

### ABSTRACT

HOTS is a very crucial thinking skill needed by teachers training students to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the teachers training students 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 teachers training student as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. The reliability test with Cronbach alpha shows a coefficient of 0.605 (reliable) for the multiple-choice and 0.61 (reliable) for the essay. The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze teachers training student' HOTS. This data can be used as the reference for developing competency improvement programs for teachers training students, for example through HOTS-oriented learning models and HOTS improvement training for teachers training student. The teacher training department can prepare learning activities that can train and empower their students' HOTS.

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

## INTRODUCTION

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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, HOTS is 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. 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 elementary school education department in science learning. What is meant by valid here is that it has been through testing by experts and empirical tests. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of elementary school education department in science learning. This instrument can be used to see the students' HOTS, so that teachers training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## **AIMS**

This study aims to develop a valid HMCEQ 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**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 students in their 2<sup>nd</sup> year in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) refers to Bloom taxonomy, consists of 1-3 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 teachers training students.

#### *c. 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 primary teacher education department who are taking a natural science course to become 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.

#### *d. 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.

## **Instrument**

### *a. 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.

*b. 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 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 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, develop, and dissemination.

**Define**

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*

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

## **Design**

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*

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

*Etc...*

---

The guidelines above were formulated in the following questions.

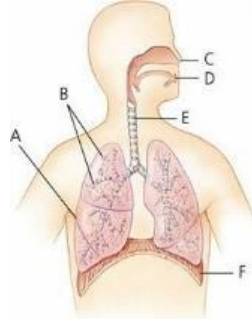
*Multiple Choice Questions*

*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 the 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 question 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

## Develop

### Validity

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*

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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*

<b>Item</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}} (0,367) = \text{valid}$*

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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***

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 indicators 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 each other. It is also suitable to questions which was 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 internal consistency reliability of items. It is an equivalent measure for dichotomous items. Meanwhile, Cronbach's alpha test is important and more useful test for internal reliability of questionnaire. It is one way concept of measuring strength of that consistency (Singh, 2017). Based on the reliability test with KR-20 on multiple-choice questions, it resulted 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 Index (DI) and Difficulty Index (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  $-1.00 \text{ logit} < b < +1.00 \text{ logit}$ , 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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

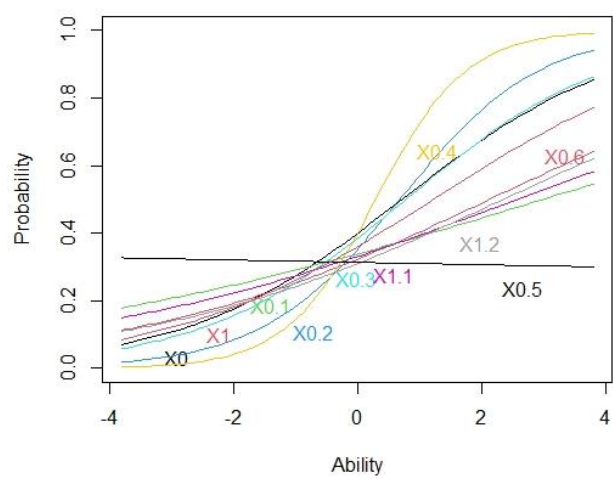
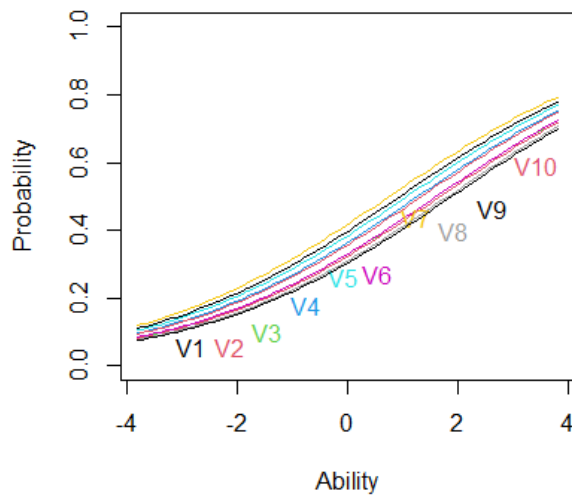


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions

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 &

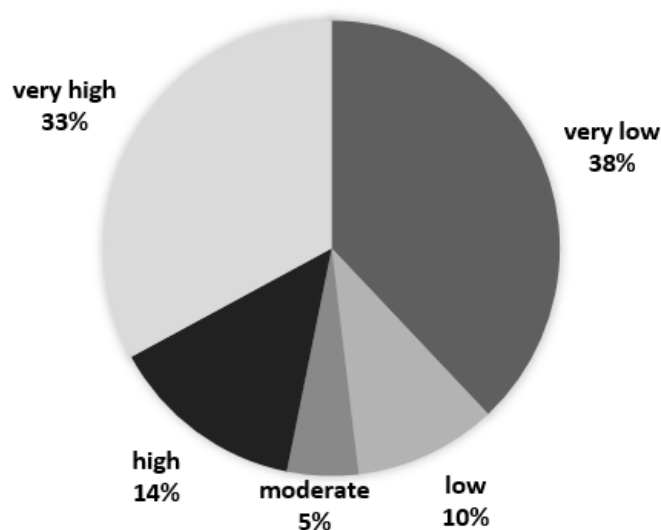


Jaleel, 2012). The effectiveness of the distractor is how well the wrong option can deceive the testees who do not know the correct answer (Herrmann-Abell et al., 2011). The more testing participants were choosing the distractor, the more it functions appropriately. A good distractor will be chosen evenly 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. Example of Distractor Revision*

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
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 teachers training students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.



*Figure 3. Analysis results of teachers training students' HOTS*

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

### **Disseminate**

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, 2002). 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 teachers training students 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 key (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 key. 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 teachers training students' 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**

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teachers training students. 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 misfit. 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 index. 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 teachers training students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *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>
- 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.
- 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., & Samuelson, 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).

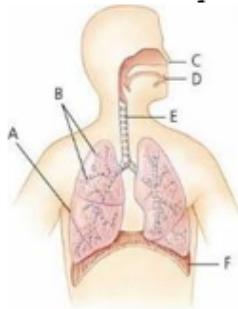
- Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*, 2(2), 155. <https://doi.org/10.21831/reid.v2i2.11029>
- Sener, N., & Tas, E. (2017). Developing Achievement Test: A Research 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\\_Validit\\_y\\_of\\_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+and+Reliability.pdf](https://www.researchgate.net/profile/Mohamed_Hammad11/post/Reliability_and_Validit_y_of_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

# REVISI ROUND 2 KE-3

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Teachers Training Students

Ika Maryani<sup>1</sup>, Zuhdan Kun Prasetyo<sup>2</sup>, Insih Wilujeng<sup>3</sup>, Siwi Purwanti<sup>4</sup>, Meita Fitriawanawati<sup>5</sup>

<sup>1</sup> Doctoral Student., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id), ORCID ID: [0000-0002-7154-2902](https://orcid.org/0000-0002-7154-2902)

<sup>2</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [zuhdan@uny.ac.id](mailto:zuhdan@uny.ac.id), ORCID ID: [0000-0001-9342-1565](https://orcid.org/0000-0001-9342-1565)

<sup>3</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [insih@uny.ac.id](mailto:insih@uny.ac.id), ORCID ID: [0000-0003-1900-7985](https://orcid.org/0000-0003-1900-7985)

<sup>4</sup> Instructor., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [siwi.purwanti@pgsd.uad.ac.id](mailto:siwi.purwanti@pgsd.uad.ac.id), ORCID ID: [0000-0002-1433-7531](https://orcid.org/0000-0002-1433-7531)

<sup>5</sup> Instructor., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [meita.fitriawanawati@pgsd.uad.ac.id](mailto:meita.fitriawanawati@pgsd.uad.ac.id), ORCID ID: [0000-0002-3748-3718](https://orcid.org/0000-0002-3748-3718)

### Correspondent

Ika Maryani, M.Pd. Universitas Ahmad Dahlan, Jl. Ki Ageng Pemanahan 19 Sorosutan, Yogyakarta-Indonesia, +6282297575204, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id).

---

### ABSTRACT

HOTS is a very crucial thinking skill needed by teachers training students to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the teachers training students 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 teachers training student as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. Based on the reliability test with KR-20 on multiple-choice questions, it resulted 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). The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze teachers training student' HOTS. This data can be used as the reference for developing competency improvement programs for teachers training students, for example through HOTS-oriented learning models and HOTS improvement training for teachers training student. The teacher training department can prepare learning activities that can train and empower their students' HOTS.

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

## **INTRODUCTION**

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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, HOTS is 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. 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 elementary school education department in science learning. What is meant by valid here is that it has been through testing by experts and empirical tests. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of elementary school education department in science learning. This instrument can be used to see the students' HOTS, so that teachers training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## **AIMS**

This study aims to develop a valid HMCEQ 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**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 students in their 2<sup>nd</sup> year in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) refers to Bloom taxonomy, consists of 1-3 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 teachers training students.

#### *c. 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 primary teacher education department who are taking a natural science course to become 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.

#### *d. 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.

## **Instrument**

### *a. 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.

*b. 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 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 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, develop, and dissemination.

**Define**

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*

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

## **Design**

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*

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

*Etc...*

---

The guidelines above were formulated in the following questions.

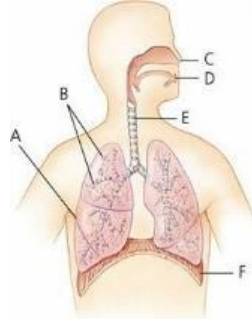
*Multiple Choice Questions*

*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 the 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 question 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

## Develop

### Validity

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*

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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*

<b>Item</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}} (0,367) = \text{valid}$*

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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***

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 indicators 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 each other. It is also suitable to questions which was 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 internal consistency reliability of items. It is an equivalent measure for dichotomous items. Meanwhile, Cronbach's alpha test is important and more useful test for internal reliability of questionnaire. It is one way concept of measuring strength of that consistency (Singh, 2017). Based on the reliability test with KR-20 on multiple-choice questions, it resulted 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 Index (DI) and Difficulty Index (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  $-1.00 \text{ logit} < b < +1.00 \text{ logit}$ , 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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

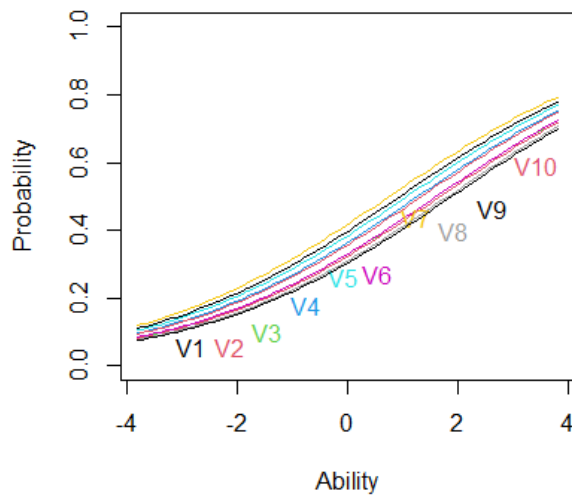


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions

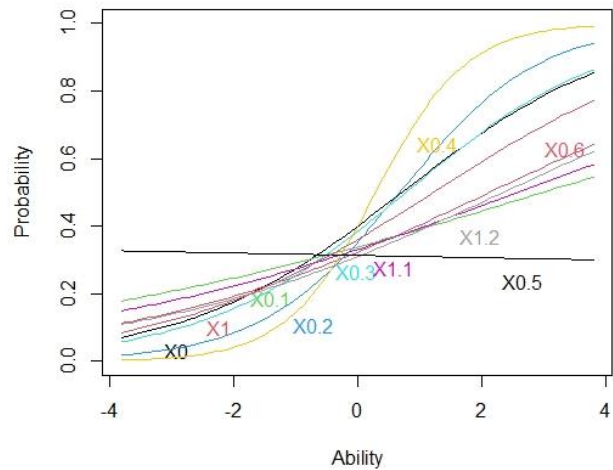


Figure 2. Result of Difficulty Index (DIF) of essay questions

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 &

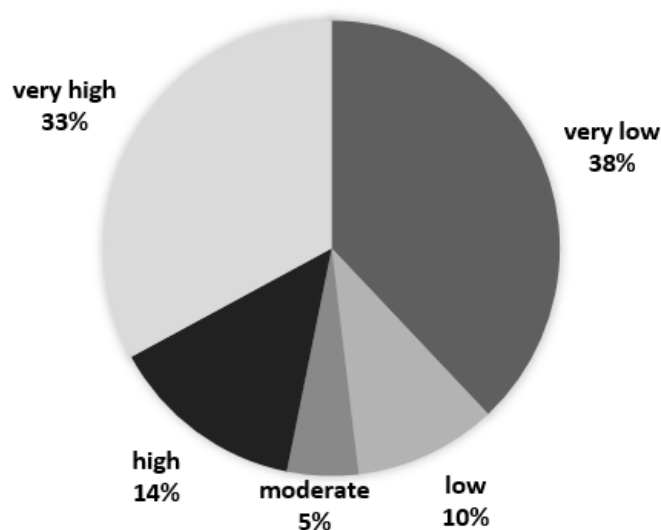


Jaleel, 2012). The effectiveness of the distractor is how well the wrong option can deceive the testees who do not know the correct answer (Herrmann-Abell et al., 2011). The more testing participants were choosing the distractor, the more it functions appropriately. A good distractor will be chosen evenly 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. Example of Distractor Revision*

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
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 teachers training students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.



*Figure 3. Analysis results of teachers training students' HOTS*

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

### **Disseminate**

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, 2002). 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 teachers training students 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 key (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 key. 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 teachers training students' 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**

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teachers training students. 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 misfit. 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 index. 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 teachers training students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *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>
- 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.
- 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., & Samuelson, 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).

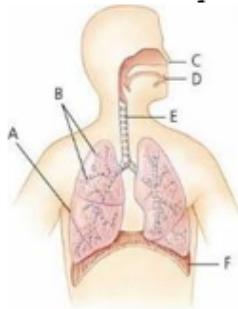
- Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*, 2(2), 155. <https://doi.org/10.21831/reid.v2i2.11029>
- Sener, N., & Tas, E. (2017). Developing Achievement Test: A Research 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\\_Validit\\_y\\_of\\_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+and+Reliability.pdf](https://www.researchgate.net/profile/Mohamed_Hammad11/post/Reliability_and_Validit_y_of_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: artikelmateri.com*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?


[← Back to Submissions](#)

Submission

Review

Copyediting

Production

## Copyediting Discussions

[Add discussion](#)

Name	From	Last Reply	Replies	Closed
<a href="#">About Proof Reading and Editing</a>	uormanci 2021-04-28 11:27 AM	ikamaryani 2021-07-16 08:48 PM	3	<input type="checkbox"/>
▶ <a href="#">article information about proofread</a>	ikamaryani 2021-06-10 04:54 AM	-	0	<input type="checkbox"/>
▶ <a href="#">Article after proofread</a>	ikamaryani 2021-11-11 04:38 PM	-	0	<input type="checkbox"/>
<a href="#">About Article Format</a>	uormanci 2021-12-22 11:45 AM	-	0	<input type="checkbox"/>
▶ <a href="#">Format of Copy Editing Article</a>	ikamaryani 2021-12-23 05:24 AM	-	0	<input type="checkbox"/>

## Copyedited

[Q Search](#)

5585	<a href="#">admin, 8-1018-Article Text-5495-1-18-20211223.pdf</a>	December 31, 2021	Article Text
------	-------------------------------------------------------------------	----------------------	--------------

99+ Compose

Mail

Inbox 4,580

Starred

Snoozed

Important

Sent

Drafts 132

Categories

More

Labels

UAD

UNIVERSITAS AHMAD DAHLAN

Q tused

27 of 43

## [tused] Editor Decision External Inbox x



**İsa DEVECİ, Assoc. Prof. Dr., Kahramanmaras Sutcu İma...**  
to me, Irma, Zuhdan, İnsih, Siwi

Wed, Feb 17, 2021, 3:47 PM

Dear Ika Maryani Maryani;

We have reached a decision regarding your submission to Journal of Turkish Science Education " **MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primary School Natural Science Teachers**". Your article was evaluated by two reviewers in Round 2. We have reached the "**Revision Required**" decision regarding your manuscript. In addition, both reviewers expressed their opinions on the fulltext (in attached files, if there is). Thus, the reviewers and editorial views pointed out some serious shortcomings and inadequacies in your work. If you think you can make these corrections, please revise your manuscript.

Our decision is to: "**Revision Required**"

**Important note:**

1. Please indicate **in color** on the fulltext corrections you have made for the opinions of both the referees and the editor (if there is).
2. Also, upload your detailed answers to the reviewers' and editorial opinions (if there is) as "**author response form**" (as a separate word file in detail) to the system.

Once again, thank you for submitting your manuscript to Journal of Turkish Science Education and we look forward to receiving your revision **in four weeks (Until 17 April 2021)**.

**Reviewer #1** (You can see the details on the attached full text)

-This article focuses on '**A Validated Instrument to Measure Higher Order Thinking Skills of Student-Teacher**'

-which is crucial to check and assess student teachers' knowledge on HOTS. However, the authors have to be consistent in addressing the respondents of the study

-as it is very confusing (student teachers OR prospective teachers or freshmen).

**Reviewer #2** (You can see the details on the attached full text)

**HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Student-Teacher**

**Abstract:**

1. Student-Teacher??? is this a prospective teacher or student and teacher???

**BUKTI REVISI  
PASCA  
ACCEPTED  
UNTUK COPY  
EDITING**

99+

Compose

Mail

Inbox 4,580

Starred

Snoozed

Important

Sent

Drafts 132

Categories

More

Labels

UAD

20 of 43

[tused] Editor Decision External Inbox x



**Assoc. Prof. Dr. İsa DEVECI** [deveciisa@gmail.com](mailto:deveciisa@gmail.com) via [tused.org](mailto:tused.org) Tue, Apr 6, 2021, 4:29 PM  
to me, Irma, Zuhdan, Insih, Siwi

Ika Maryani, Irma Rifda Syahada, Zuhdan Kun Prasetyo, Insih Wilujeng, Siwi Purwanti:

We have reached a decision regarding your submission to Journal of Turkish Science Education, "MCEQ (Multiple Choice and Essay Questions): A Validated Instrument for Measuring Higher Order Thinking Skills of Pre-service Primar School Natural Science Teachers".

Our decision is to: Accept Submission

Assoc. Prof. Dr. İsa DEVECI  
[deveciisa@gmail.com](mailto:deveciisa@gmail.com)

**BUKTI ACCEPT**

[Journal of Turkish Science Education](#)

One attachment • Scanned by Gmail



Reply

Reply all

Forward

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of Teachers Training Students

Ika Maryani<sup>1</sup>, Zuhdan Kun Prasetyo<sup>2</sup>, Insih Wilujeng<sup>3</sup>, Siwi Purwanti<sup>4</sup>, Meita Fitriawanawati<sup>5</sup>

<sup>1</sup> Doctoral Student., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id), ORCID ID: [0000-0002-7154-2902](https://orcid.org/0000-0002-7154-2902)

<sup>2</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [zuhdan@uny.ac.id](mailto:zuhdan@uny.ac.id), ORCID ID: [0000-0001-9342-1565](https://orcid.org/0000-0001-9342-1565)

<sup>3</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [insih@uny.ac.id](mailto:insih@uny.ac.id), ORCID ID: [0000-0003-1900-7985](https://orcid.org/0000-0003-1900-7985)

<sup>4</sup> Instructor., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [siwi.purwanti@pgsd.uad.ac.id](mailto:siwi.purwanti@pgsd.uad.ac.id), ORCID ID: [0000-0002-1433-7531](https://orcid.org/0000-0002-1433-7531)

<sup>5</sup> Instructor., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [meita.fitriawanawati@pgsd.uad.ac.id](mailto:meita.fitriawanawati@pgsd.uad.ac.id), ORCID ID: [0000-0002-3748-3718](https://orcid.org/0000-0002-3748-3718)

### Correspondent

Ika Maryani, M.Pd. Universits Ahmad Dahlan, Jl. Ki Ageng Pemanahan 19 Sorosutan, Yogyakarta-Indonesia, +6282297575204, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id).

---

### ABSTRACT

HOTS is a very crucial thinking skill needed by teachers training students to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the teachers training students 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 teachers training student as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. Based on the reliability test with KR-20 on multiple-choice questions, it resulted 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). The discrimination index showed discarded, sufficient, good, and very good. The 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze teachers training student' HOTS. This data can be used as the reference for developing competency improvement programs for teachers training students, for example through HOTS-oriented learning models and HOTS improvement training for teachers training student. The teacher training department can prepare learning activities that can train and empower their students' HOTS.

**Keywords:** *HMCEQ, instrument, higher-order thinking skills.*

---

## INTRODUCTION

21<sup>st</sup>-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 & Maryani (2017) explained that the institute has an important task to prepare professional teachers. Teachers are professional occupations that 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, HOTS is 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. The lesson plan must be adjusted to the demands of the curriculum and must allow students to think and analyze critically (Nursalam & Rasyid, 2016). 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 higher-order thinking skills (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 is a mental process that requires 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 (S. C. 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), the development of HOTS-Based Mathematical E-Quiz (Electronic Quiz) Assessment Instrument for Grade 5 of primary school (Nur Aini & Sulistyani, 2019). Besides, Syafri Ahmad, Kenedi, & Masniladevi (2018) has 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. Syafri 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) (S. 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 elementary school education department in science learning. What is meant by valid here is that it has been through testing by experts and empirical tests. Therefore, it is urgent to develop a valid instrument to measure the students' HOTS of elementary school education department in science learning. This instrument can be used to see the students' HOTS, so that teachers training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## **AIMS**

This study aims to develop a valid HMCEQ 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**

### ***Participant***

The research subjects consist of subjects for testing and subjects for implementation. In the development step, 81 students in their 2<sup>nd</sup> year in primary teacher education were selected to participate. In contrast, in the disseminate step, 75 students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

### ***Development Framework***

This research and development 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 was employed.

#### ***a. 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.

#### ***b. 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 HOTS indicator (analysis, evaluation, creation) refers to Bloom taxonomy, consists of 1-3 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 teachers training students.

#### *c. 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 primary teacher education department who are taking a natural science course to become 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.

#### *d. 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.

## **Instrument**

### *a. 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.

#### *b. 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 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 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, develop, and dissemination.

### **Define**

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*

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

## **Design**

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*

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

*Etc...*

---

The guidelines above were formulated in the following questions.

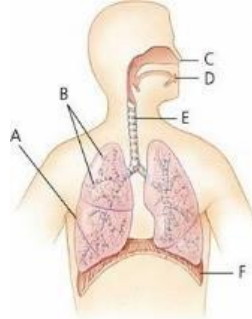
*Multiple Choice Questions*

*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 the 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 question 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

## Develop

### Validity

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*

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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*

<b>Item</b>	<b>R<sub>value</sub></b>	<b>Criteria</b>
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

*Note:  $R > R_{\text{table}} (0,367) = \text{valid}$*

Based on Tables 4 and 5, 3 items in multiple choices questions are misfit and 7 items are fit, whereas 2 items in essay question 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
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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***

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 indicators 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 each other. It is also suitable to questions which was 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 internal consistency reliability of items. It is an equivalent measure for dichotomous items. Meanwhile, Cronbach's alpha test is important and more useful test for internal reliability of questionnaire. It is one way concept of measuring strength of that consistency (Singh, 2017). Based on the reliability test with KR-20 on multiple-choice questions, it resulted 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 Index (DI) and Difficulty Index (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  $-1.00 \text{ logit} < b < +1.00 \text{ logit}$ , 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.  $D \geq 0.4$  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  $D \leq 0$ , 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.

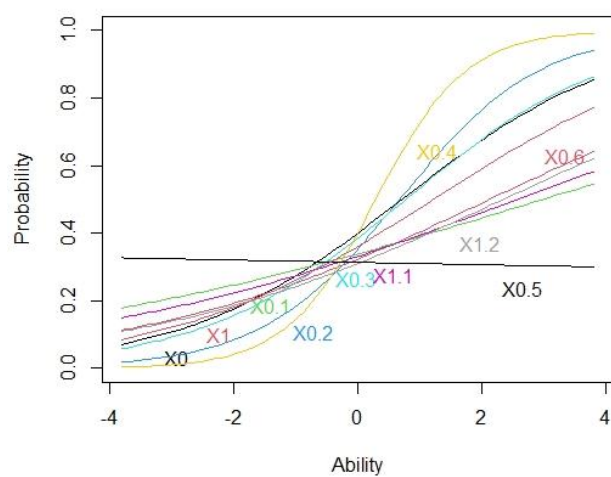
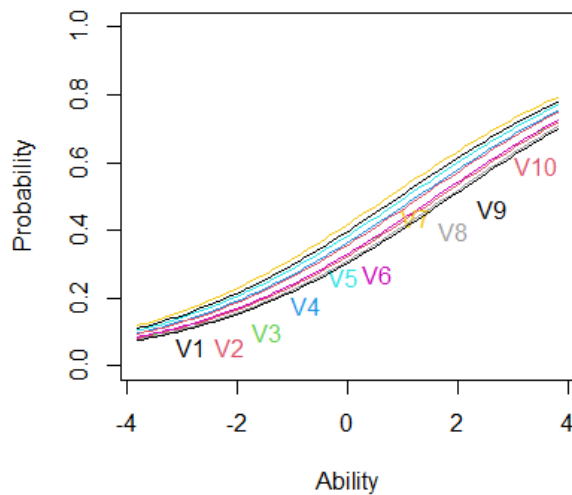


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions

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)

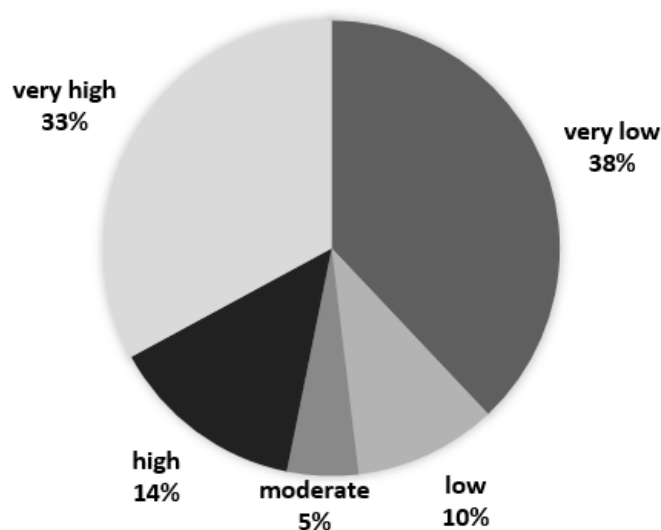
In multiple-choice questions, there is an option that functions as a distractor. The distractor works effectively if it is chosen by at least 5% of all testing participants (Hingorjo &

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. Example of Distractor Revision*

Number of Question Item	Answer Choices	Distractors		Purpose of Revision
		Before Revision	After Revision	
<b>Question set A</b>				
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 teachers training students taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.



*Figure 3. Analysis results of teachers training students' HOTS*

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

### **Disseminate**

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, 2002). 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 teachers training students 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 key (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 key. 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 teachers training students' 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**

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of teachers training students. 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 misfit. 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 index. 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 teachers training students' 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.

## REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnampereuma, 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. (2002). *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>
- 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.
- 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., & Samuelson, 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).

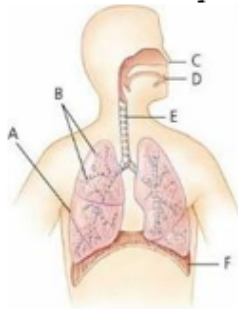
- Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *Research and Evaluation in Education*, 2(2), 155. <https://doi.org/10.21831/reid.v2i2.11029>
- Sener, N., & Tas, E. (2017). Developing Achievement Test: A Research 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\\_Validit\\_y\\_of\\_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+and+Reliability.pdf](https://www.researchgate.net/profile/Mohamed_Hammad11/post/Reliability_and_Validit_y_of_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+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.

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

*A2. Look at the picture below!*



*A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...*

*Source: [artikelmateri.com](http://artikelmateri.com)*

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?



tused



# BUKTI PROSES PRODUCTION

10 of 43

- 99+
- Mail
  - Compose
  - Inbox** 4,580
  - Starred
  - Snoozed
  - Important
  - Sent
  - Drafts** 132
  - Categories
  - More
- Chat
- Spaces
- Meet
- Labels
  - UAD

[tused] Editor Decision External Inbox x



**Bugra Ulger** b.bugra84@gmail.com via tused.org to me, Insih, Siwi

Fri, Dec 31, 2021, 6:23 PM

Ika Maryani, Zuhdan Kun Prasetyo, Insih Wilujeng, Siwi Purwanti, Meita Fitriawanati:

The editing of your submission, "HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers," is complete. We are now sending it to production.

Submission URL: <http://tused.org/index.php/tused/authorDashboard/submission/1018>

Bugra Ulger  
[b.bugra84@gmail.com](mailto:b.bugra84@gmail.com)

---

[Journal of Turkish Science Education](#)

- Reply
- Reply all
- Forward

**Journal of Turkish Science Education**

<http://www.tused.org>

© ISSN: 1304-6020

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

Ika Maryani<sup>1</sup>, Zuhdan Kun Prasetyo<sup>2</sup>, Insih Wilujeng<sup>3</sup>, Siwi Purwanti<sup>4</sup>, Meita Fitriawanawati<sup>5</sup>

<sup>1</sup> Doctoral Student of Educational Science Department, Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, ORCID ID: 0000-0002-7154-2902

<sup>2</sup> Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, ORCID ID: 0000-0001-9342-1565

<sup>3</sup> Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, ORCID ID: 0000-0003-1900-7985

<sup>4</sup> Department of Elementary School Teacher Education, Universitas Ahmad Dahlan, Yogyakarta-Indonesia, ORCID ID: 0000-0002-1433-7531

<sup>5</sup> 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**

21<sup>st</sup>-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 become 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 2<sup>nd</sup> year in primary teacher education were selected to participate. In contrast, in the dissemination step, 75 students in their 1<sup>st</sup> 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.

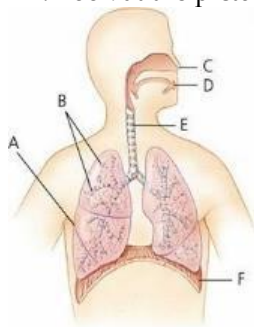
Multiple Choice Questions

A1. The lungs function to transport oxygen from the air into the bloodstream.

It indicates that the lungs...

- have a wide surface
- have an elastic surface
- are rich in capillary
- are protected by a pleural membrane
- 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
- B
- C dan D
- C dan D
- 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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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 <sub>value</sub>	Criteria
B1	0,548	Valid
B2	0.286	Invalid
B3	0,743	Valid
B4	0,203	Invalid
B5	0,470	Valid

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

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

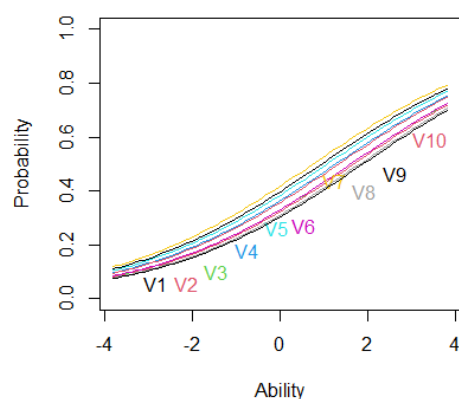
The reliability test of HMCEQ is related to the accuracy of the test results (Heale & Twycross, 2015). Reliability is used to measure the consistency of a test. It is used to test the consistency of the question items when the test was taken repeatedly by the same object (Bajpai & Bajpai, 2014; Beck 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.  $D \geq 0.4$  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  $D \leq 0$ , 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*

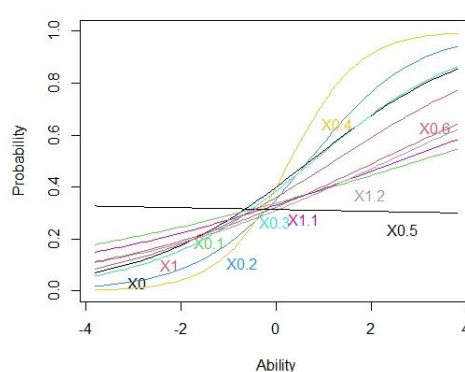


Figure 2 shows that the order of the difficulty index for multiple-choice questions from the easiest to the most difficult is V7-V1-V5-V4-V2-V6-V10-V8-V9, while for essay questions from easy to difficult are X1-X2-X4-X3-X5. The difficulty index and discriminant index data are shown in Table 7.

**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 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

**Figure 3**

*Analysis results of prospective teachers' HOTS*

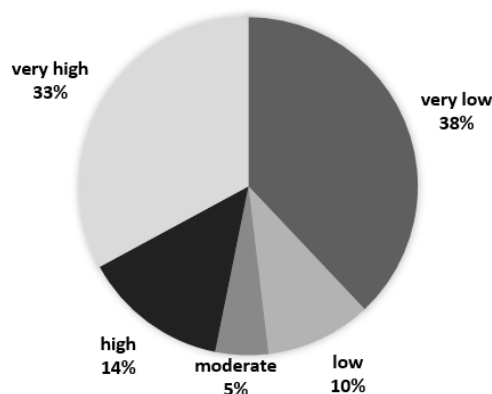


Figure 3 shows that most prospective teachers have very low HOTS (38%) and very high HOTS (33), while 14% have high HOTS, 10% low, and 5% moderate.

### 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 (Thanerananon 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  $1\text{NFD} < 2\text{NFD} < 3\text{NFD}$ . 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., & Ponnamparuma, 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.000000000000016>
- 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. *WMI 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., Widiastuti, 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., & Samuelson, 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](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+validity+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>



# BUKTI PROSES PROOFREAD

## HMCEQ (HOTS Multiple Choice and Essay Questions): A Validated Instrument to Measure Higher Order Thinking Skills of ~~Prospective Teachers~~ ~~Training Students~~

**Commented [A1]:** Please re-consider using abbreviations in the title

**Commented [A2]:** Or preservice teachers, or teacher candidates. But I suggest the term prospective.

Ika Maryani<sup>1</sup>, Zuhdan Kun Prasetyo<sup>2</sup>, Insih Wilujeng<sup>3</sup>, Siwi Purwanti<sup>4</sup>, Meita Fitriawanawati<sup>5</sup>

<sup>1</sup> Doctoral Student., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id), ORCID ID: [0000-0002-7154-2902](https://orcid.org/0000-0002-7154-2902)

<sup>2</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [zuhdan@uny.ac.id](mailto:zuhdan@uny.ac.id), ORCID ID: [0000-0001-9342-1565](https://orcid.org/0000-0001-9342-1565)

<sup>3</sup> Prof. Dr., Universitas Negeri Yogyakarta, Yogyakarta-Indonesia, [insih@uny.ac.id](mailto:insih@uny.ac.id), ORCID ID: [0000-0003-1900-7985](https://orcid.org/0000-0003-1900-7985)

<sup>4</sup> Instructor., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [siwi.purwanti@pgsd.uad.ac.id](mailto:siwi.purwanti@pgsd.uad.ac.id), ORCID ID: [0000-0002-1433-7531](https://orcid.org/0000-0002-1433-7531)

<sup>5</sup> Instructor., Universitas Ahmad Dahlan, Yogyakarta-Indonesia, [meita.fitriawanawati@pgsd.uad.ac.id](mailto:meita.fitriawanawati@pgsd.uad.ac.id), ORCID ID: [0000-0002-3748-3718](https://orcid.org/0000-0002-3748-3718)

### Correspondent

Ika Maryani, M.Pd. Universitas Ahmad Dahlan, Jl. Ki Ageng Pemanahan 19 Sorosutan, Yogyakarta-Indonesia, +6282297575204, [ika.maryani@pgsd.uad.ac.id](mailto:ika.maryani@pgsd.uad.ac.id).

## ABSTRACT

HOTS ~~is~~ ~~are~~ very crucial thinking skills needed by teachers ~~to training~~ students to develop 21st-century learning. This study aimed to develop HMCEQ to measure the higher-order thinking skills of the ~~teachers-training student~~ ~~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~~ ~~training-student~~ as the test subjects. The assessment of instrument quality by experts showed that the average score of the question quality was 81.16 (very good). This research succeeded in developing HMCEQ which consisted of 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 (number 3, 9, and 10) were classified as a misfit, while the 2 essay questions are invalid and the other (3 questions) as valid. 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). 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. The implementation of the test showed that 33% of the questions were very high, 14% were high, 5% were moderate, 10% were low, and 38% were very low. This instrument can be used to analyze ~~teachers-training student~~ ~~prospective teachers~~' HOTS ~~higher-order thinking skills~~. This data can be used as the reference for developing competency improvement programs for ~~teachers-training student~~ ~~prospective teachers~~, for example through ~~higher-order thinking skills~~ ~~HOTS~~-oriented learning models and ~~higher-order thinking skills~~ ~~HOTS~~ improvement training for ~~teachers-training student~~ ~~prospective teachers~~. The teacher training department can prepare learning activities that can train and empower their students' ~~higher-order thinking skills~~ ~~HOTS~~.

**Commented [A3]:** Please avoid using abbreviations in the abstract.

**Commented [A4]:** Please avoid using abbreviations in the abstract.

**Keywords:** ~~HMCEQ~~, instrument, higher-order thinking skills.

**Commented [A5]:** this abbreviation will not benefit you in indexing.

## INTRODUCTION

21<sup>st</sup>-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 professional ~~occupations that-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) ~~is-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-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-21st-~~century skill of ~~teachers-training-student~~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 ~~teachers-training-student~~prospective teachers' ~~21st-21st-~~century skills in science learning ~~was-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 ~~higher-order thinking skills (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 ~~is-are~~ mental processes ~~es~~ that requires 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 (~~S.-C.~~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, ~~Syafri-Ahmad, Kenedi, & Masniladevi et al.~~ (2018) ~~has-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. ~~Syafri-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) (~~S.-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 empirical ~~testsly~~. 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 teachers' training department can use this data to develop HOTS training and empowerment programs and recommend appropriate learning models to improve HOTS.

## AIMS

This study aims to develop a valid HMCEQ 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.

**Commented [A6]:** Scale? Measurement tool?

## METHODS

### Research Design

#### Participants

The research subjects-participants consist of subjects for testing and subjects for implementation. In the development step, 81 students in their 2<sup>nd</sup> year in primary teacher education were selected to participate. In contrast, in the disseminate-dissemination step, 75 students in their 1<sup>st</sup> 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 refers to (Creswell, 2012). The number of samples has met the criteria of sample size in descriptive research.

**Commented [A7]:** Please provide detailed and extensive information about the method you used and the reasons for you choosing it.

**Formatted:** Font: Italic

**Formatted:** Font: Italic, Turkish (Türkiye)

#### Development Framework

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.

**Commented [A8]:** Please provide evidence of sample selection method.

**Commented [A9]:** ?

**Commented [A10]:** Provide evidence of the number please.

**Commented [A11]:** Please check the terminology. I suggest scale development maybe.

#### *a. 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.

*b. 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 HOTS indicator (analysis, evaluation, creation) refers to Bloom taxonomy, consists of 1-3 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 ~~teachers training student~~prospective teachers.

*c. 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 become 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.

*d. Disseminate*

**Commented [A12]:** Each indicator should consist of more than 2 items. Please support your choice here.

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.

## **Instrument**

### *a. 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.

### *b. 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

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*

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

### Design

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	HOTS 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
<i>Etc...</i>					

**Commented [A13]:** I have strong doubts about the high-levels of the questions below. Please elaborate on the levels and provide evidence to the level of the questions.

The guidelines above were formulated in the following questions.

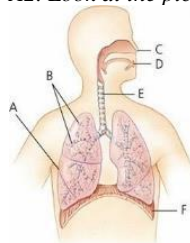
*Multiple Choice Questions*

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 ~~the 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**

**Validity**

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*

No	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
<b>Average</b>		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 <sup>2</sup>	Pr (> X <sup>2</sup> )	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 <sub>value</sub>	Criteria
<b>B1</b>	0,548	Valid
<b>B2</b>	0.286	<b>Invalid</b>
<b>B3</b>	0,743	Valid
<b>B4</b>	0,203	<b>Invalid</b>
<b>B5</b>	0,470	Valid

Note:  $R > R_{table} (0,367) = \text{valid}$

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.

Commented [A14]: What precautions were taken for misfit questions?

Table 6. Revision of Invalid Questions

Question	Before revision			After revision		
	Indicators	Questions	Cognitive Level	Indicators	Questions	Cognitive Level
<b>B2</b>	A statement is presented, students can clarify why the lungs are not injured despite experiencing friction	Inspiration and expiration make the 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
<b>B4</b>	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 bear it. 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**

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 indicators 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 each the other. It is also suitable for questions which that was-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-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).

**Commented [A15]:** Reliability below .70 can be considered as not enough. Please provide evidence that it is.

### **Discrimination Index (DI) and Difficulty Index (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  $-1.00 \text{ logit} < b < +1.00 \text{ logit}$ , and items are categorized as difficult if the value of b approaches +2.00 logit. Furthermore, items with a value of  $b > +2.00 \text{ logit}$  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.  $D \geq 0.4$  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  $D \leq 0.20$  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.

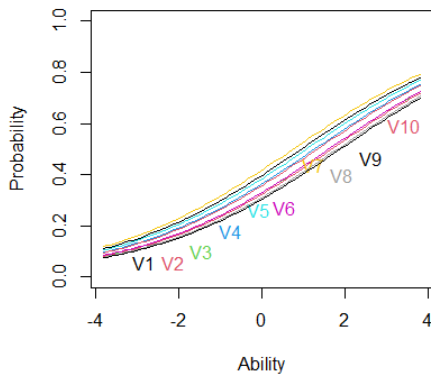


Figure 1. Result of Difficulty Index (DIF) of multiple-choice questions

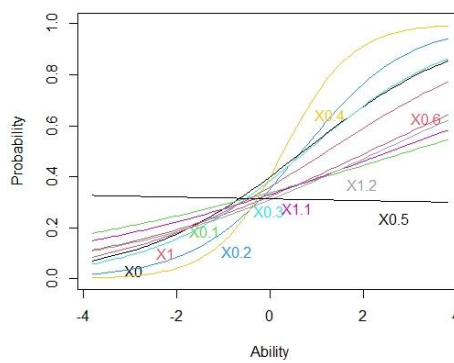


Figure 2. Result of Difficulty Index (DIF) of essay questions

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

Commented [A16]: It is not clear how these result support the data in the figures.

### 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	
<b>Question set A</b>				
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 ~~teachers training student~~ prospective teachers taking the Natural Science course in the 5<sup>th</sup> semester of elementary school teacher education. The results of the analysis are presented in Figure 3.

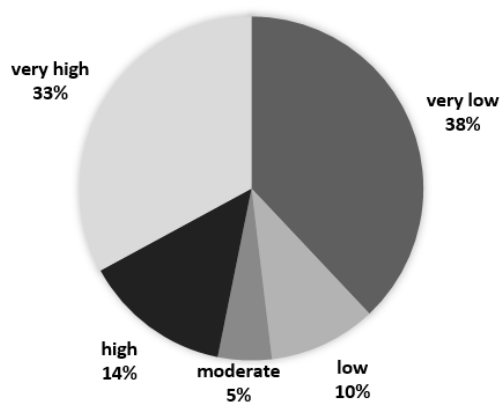


Figure 3. Analysis results of ~~teachers training student~~prospective teachers~~teachers~~' HOTS

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

### Disseminate

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, 2002). 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 ~~teachers training student~~prospective teachers~~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 ~~teachers training student~~ 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

This research has succeeded in producing HMCEQ on natural science to measure the higher-order thinking skills of ~~teachers training student~~ 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 ~~teachers training student~~ ~~prospective teachers~~ ~~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.

#### REFERENCE

- Abdulghani, H., Ahmad, F., Aldrees, A., Khalil, M., & Ponnamparuma, 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., 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>~~
- 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 S., 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. (2002). *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.000000000000016>
- 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>
- Chebbi, 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., Widiastuti, 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/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+and+Reliability.pdf](https://www.researchgate.net/profile/Mohamed_Hammad11/post/Reliability_and_Validity_of_Scales/attachment/5a150ca24cde26c48ab5d328/AS:563368521547776@1511328930210/download/2017+COMMON+PROCEDURES+FOR+DEVELOPMENT%2C+VALIDITY+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*

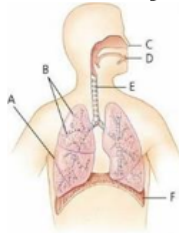
*Education*, 15(6), 725–737.

- 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>

## APPENDICES

### A sample item for the Multiple-Choice Question (C5)

A2. Look at the picture below!



A2. In the respiratory system the organ that becomes the location where oxygen exchanges with carbon dioxide is indicated by letter...

Source: [artikelmateri.com](http://artikelmateri.com)

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

### A sample item for the Essay Question (C4)

A passive smoker is people who inhale cigarette smoke from people who smoking or people who are exposed to secondhand smoke from smoke excluded by a passive smoker. He/she has a higher risk compared to an active smoker. Even the dangers that must be borne by a passive smoker three to five times the danger of an active smoker. Why does it happen?

**Commented [A17]:** The whole tool will be expected here. These questions are already in the text, so these here becomes redundant. Also, rubrics and the answer key should be given here.



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

Research Article

**Ika Maryani**

Universitas Ahmad Dahlan

 <https://orcid.org/>**Zuhdan Kun Prasetyo****Insih Wilujeng**

Universitas Negeri Yogyakarta

**Siwi Purwanti**

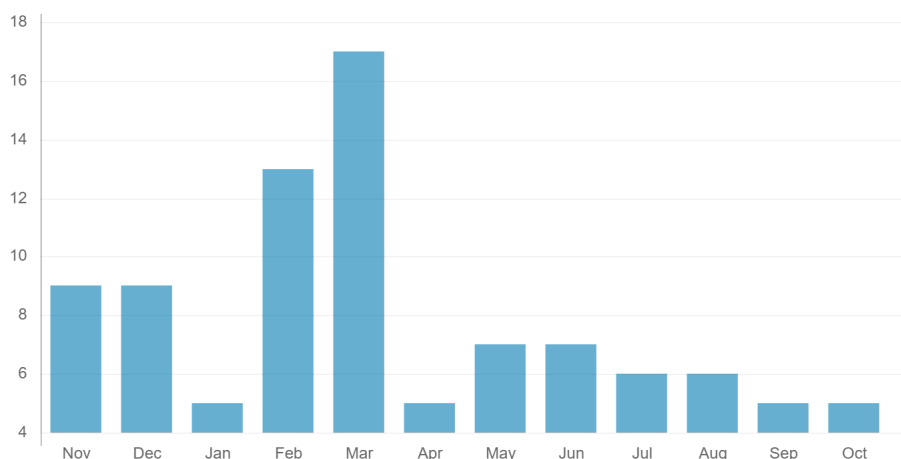
Universitas Ahmad Dahlan

**Meita Fitriawanati****Keywords:** Higher-order thinking skill, multiple choice, essay questions, instrument

### 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.

### Downloads

 PDF1

Published

2021-12-31

Issue

**Most read articles by the same author(s)**

- Ahmad Syawaludin, Zuhdan Kun Prasetyo, Cipi Safruddin Abdul Jabar, Heri Retnawati, [The Effect of Project-based Learning Model and Online Learning Setting to Analytical Skills of Discovery Learning, Interactive Demonstrations, and Inquiry Lessons on the Pre-Service Elementary Teachers](#), *Journal of Turkish Science Education: Vol. 19 No. 2 (2022): The Journal of Turkish Science Education*
- Nur ATIKOH, Zuhdan Kun PRASETYO, [The Effect of Picture Storybook Based on Scientific Approach through Inquiry Method toward Student's Inference Skill](#), *Journal of Turkish Science Education: Vol. 15 No. Special (2018): Special Issue*



[Make a Submission](#)

**Information**

[For Readers](#)

[For Authors](#)

[For Librarians](#)



**Readers**

- [Read Articles](#)
- [View All Issues](#)

**Authors**

- [Author Information Pack](#)
- [Submit Paper](#)
- [Contact](#)

**Reviewers**

- [Become Reviewer](#)
- [Reviewer Guidelines](#)
- [Policies](#)

**Editors**

- [Editors](#)
- [Ethics](#)

Platform & workflow by  
**OJS / PKP**