*International Journal of Instruction* *October 2020 ● Vol.13, No.4*
*e-ISSN: 1308-1470 ● www.e-iji.net* *p-ISSN: 1694-609X*
*pp. 833-848*

Received: 24/09/2019
Revision: 17/05/2020
Accepted: 07/06/2020
OnlineFirst:29/08/2020

# The Instrument Development to Measure Higher-Order Thinking Skills for Pre-Service Biology Teacher

**Candra Utama**
The doctoral program of Education Science of Sebelas Maret University, Surakarta, Indonesia, *amaacandraa@student.uns.ac.id*

**Sajidan**
Faculty of Teacher Training and Education of Sebelas Maret University, Surakarta, Indonesia, *sajidan@fkip.uns.ac.id*

**Joko Nurkamto**
Faculty of Teacher Training and Education of Sebelas Maret University, Surakarta, Indonesia, *jokonurkamto@gmail.com*

**Wiranto**
Faculty of Teacher Training and Education of Sebelas Maret University, Surakarta, Indonesia, *wiranto@staff.uns.ac.id*

The 21st century demands mastery of various skills, and one of them is a higher-order thinking skill. Specialized instruments are required to measure these skills to be aware of their level of achievement. This research is classified as development research to create tools capable of measuring higher-order thinking skills of pre-service biology teachers. The development procedure is consisting of four stages, including instrument design, instrument trial, determination of validity, and determination of reliability. The instrument developed consisted of 12 items, which were then validated by experts. Trial stage involving 110 pre-service biology teachers in the Biology Education Program of eleven March University. The tested instrument is analysed to determine its validity and reliability. The results of the construct validation show model classified as fit according to the indicator of each instrument. The Results of the items analysis showed all items have an excellent discriminating power index. Also, each piece also has a difficulty index that is classified as a medium. Based on the results, the instrument developed has the qualifications to measure the higher-order thinking skills (transfer of knowledge) of pre-service biology teachers.

Keywords: construct validity, confirmatory factor analysis, taxonomy Bloom, level of difficulty, discrimination power index

**Citation:** Utama, C., Sajidan, Nurkamto, J., & Wiranto. (2020). The Instrument Development to Measure Higher-Order Thinking Skills for Pre-Service Biology Teacher. *International Journal of Instruction*, 13(4), 833-848. https://doi.org/10.29333/iji.2020.13451a

## INTRODUCTION

The object of the research most frequently used by many researchers is about 21st-century skills. Some characteristics of 21st-century skills are skillful in working independently and in groups, having creative innovation, having the critical selection and sorting of information, and having the necessary knowledge to be applied throughout life (Afandi & Sajidan, 2017). The same thing is emphasized by P21 (Partnership For 21st Century Skills) (2012) that 21st-century skills are critical, creative, communicative, and collaborative thinking skills, the abilities to solve problems, to work together, to utilize information technology, to renew, to be aware of globalization, be mindful of the environment and several other competencies. Therefore, all types and levels of education must be able to encourage students to master 21st-century skills which are not only challenges and demands but also opportunities for those who can use them.

Demands for mastering 21st-century skills will have an impact on the education and learning system. Education and schools will gradually change their roles, considering that several developed countries have moved from an industrial-based to an economic information-based educational system, and even now based on information technology. Such conditions require relevant parties, especially in the field of education to design curriculum and educational objectives tailored to the development of science, technology and information systems so that students are ready to face the various challenges in the future (Kong et al., 2014; van Laar, van Deursen, van Dijk, & de Haan, 2017).

The higher education system with its faculties which produces pre-service teachers has a greater responsibility than the others (Shuler, Winters, & West, 2013) because the teacher becomes the main gate in conveying not only material information but also science with the latest developments (Mazalah et al., 2016). Therefore, they must be open to all kinds of changes and be able to adapt to the development of increasingly advanced technology and information systems (García-Martín & García-Sánchez, 2017; Uerz, Volman, & Kral, 2018; Utama, Sajidan, Nurkamto, & Wiranto, 2019). This increasingly open and connected world causes science to be easily obtained from any part of the world (Bilyalova, Salimova, & Zelenina, 2020). Sources of science that can be obtained from anywhere require students to have the ability to access, use, and utilize them. Likewise, universities must be able to facilitate and empower their human resources, both students and educators, so that they can quickly adapt to the latest developments in science and technology (Bilyalova et al., 2020; Jeffrey, Milne, Suddaby, & Higgins, 2014).

These increasingly complex, global, and interrelated changes must be faced and responded wisely by both students and universities. Competencies needed by the world of work no longer refer to educational qualifications because competence is not always directly proportional to educational qualifications (Oey-Gardiner et al., 2017). Higher education has a vital role in this kind of change because it has a significant role in creating college graduates. If higher education is unable to prepare for change, there will be a mismatch between graduates and the needs of the world of work and society.

One way for higher education to prepare graduates who are ready to face these changes is to develop a higher education curriculum by the demands of the 21st-century. HOTS (transfer of knowledge) is the estuary of different types of skills. HOTS are part of 21st-century skills that have many definitions (Saido, Siraj, & Nordin, 2015). However, HOTS that can be inserted in the development of higher education curriculum as a transfer of knowledge of the cognitive domain of analysis, evaluation, and creation of the Bloom's taxonomy revised by Anderson (Masigno, 2014). Through a curriculum based on HOTS (transfer of knowledge), higher education can prepare themselves for change.

HOTS can be taught through various learning models and methods such as laboratory practicum; (Liu, Wu, Wong, Lien, & Chao, 2017); flipped classroom and e-learning (Lee, Lim, & Kim, 2017); collaborative discussion (Afandi & Sajidan, 2017); blended learning (Fryer & Bovee, 2018); and mobile learning (Sung, Chang, & Liu, 2016).

Higher-order thinking is not new, but it is still rarely used in learning. According to Heong et al., (2011), HOTS is one part of creative and critical thinking. Meanwhile, Masigno (2014) conveyed that the concept of HOTS comes from the cognitive domain of Bloom's taxonomy. Based on the results of the study on the cognitive area in Bloom's taxonomy, there are six cognitive levels. The three fundamental skills include the abilities to remember, understand and apply categorized as lower-order thinking (LOT) and the next three levels that have been revised by (Anderson & Krathwohl, 2001) include the abilities to analyze, evaluate and create categorized as higher-order thinking (HOT).

Based on theoretical reviews and relevant studies, the research questions are formulated as follows: (1) How is the result of the construct validity of the HOTS instrument developed? (2) Does the HOTS assessment instrument consist of feasible items based on the discriminating power and difficulty level?

The basic principle of developing the HOTS instrument as a transfer of knowledge in learning is the learning achievement to be achieved referring to the highest cognitive domains of Bloom's Taxonomy (analysis, evaluation, and creation). To assess HOTS, an instrument that involves critical-thinking skills, problem-solving skills, and creativity are needed so that the competency-based device related to learning is necessary (Yanto, Subali, & Suyanto, 2019). The teacher must plan well and involve students in learning activities that can encourage and develop HOTS. The instrument developed used problems that require the use of knowledge and skills in new situations by providing a descriptive statement (Prayitno, Suciati, & Titikusumawati, 2018).

Many experts have examined the importance of HOTS like Tan & Halili (2015), confirming that HOTS is a part of 21st-century skills that is crucial for future generations. They also compare HOTS and LOTS, the development of which requires the same amount of time, cost, and mind. HOTS has an essential role in the problem-solving process (Moore & Rubbo, 2012). The study of Yee, Lai, Tee, & Mohamad (2016) entitled "The Role of Higher-Order Thinking Skills in Green Skill Development" proves that HOTS can create new competencies appropriate to industrial developments

such as the revolution of industry 4.0. One of the characteristics of HOTS-based instruments is that each student is encouraged to participate in making decisions about various scientific problems, so they need to develop reasoning and logical abilities based on scientific knowledge (Jeong, Kim, Chae, & Kim, 2014). Also, increasing HOTS has a positive effect on learning practices (Yanto et al., 2019).

This research aims to develop an instrument that can measure the HOTS (transfer of knowledge) pre-service biology teachers in the general biology course and to obtain the characteristics of HOTS assessment that include the aspects of analysis, evaluation, and creation.

## CONTEXT AND REVIEW OF LITERATURE

### Higher-Order Thinking Skills (Transfer of Knowledge)

HOTS, as a transfer of knowledge, refers to the revised Bloom's educational taxonomy. Anderson & Krathwohl (2001) revised Bloom's taxonomy by creating a new educational taxonomy structure that is the dimension of knowledge and cognitive processes. The knowledge dimension consists of factual, conceptual, procedural, and metacognitive knowledge. The aspects of cognitive processes include remembering, understanding, applying, analyzing, evaluating, and creating.

Factual knowledge is the knowledge about several separate components that have their characteristics, which are still in the form of pure information (Zohar & Cohen, 2016). Factual knowledge consists of several specific and individual components or elements (Nguyễn & Nguyễn, 2017), inversely proportional to conceptual knowledge. Conceptual knowledge is more complex and organized compared to factual knowledge. The examples of conceptual knowledge are classification and categories, principles and generalizations, and model theories and structures (Anderson & Krathwohl, 2001).

Procedural knowledge is the knowledge about the stages of doing something (Piaget & Kohler, 2014), such as skills and algorithms, techniques and methods, and determining when to do something (Liu et al., 2017). Metacognitive knowledge is the knowledge about self-awareness (Nguyễn & Nguyễn, 2017) including strategic knowledge and knowledge of cognitive processes, including contextual and conditional knowledge and self-knowledge (Anderson & Krathwohl, 2001).

The cognitive process dimension has six levels and several indicators based on Bloom's taxonomy, revised by Anderson & Krathwohl (2001). The first level is remembering (C1), taking knowledge from long-term memory. The indicators of remembering are recognizing and recalling. The second level is understanding (C2), constructing meaning from learning the material, including what is said, written, and drawn by the educator. The indicators are interpreting, modeling, classifying, summarizing, concluding, comparing, and explaining. The third level is applying (C3), applying or using a procedure in certain circumstances. The indicators include executing and implementing. The fourth level is analyzing (C4), dividing the material into several parts and determining the relationship of each piece, between parts and the whole. The indicators are differentiating, organizing, and attributing. The fifth level is evaluating (C5), making

decisions based on standards or criteria such as checking and criticizing. The sixth level is creating (C6), which combines several parts to form something new and original. The indicators are formulating, planning, and producing.

In this study, HOTS as a transfer of knowledge is part of the highest levels of higher-order thinking skills in revised Bloom's cognitive taxonomy, which include analytical, evaluation, and creative skills in cognitive processes. The instruments are arranged based on the aspects, descriptions, and indicators of HOTS that are reconstructed and used to organize the items. Table 1 below presents the constructs of HOTS.

Table 1
HOTS Construct

| Aspect | Description | Indicator |
|---|---|---|
| Analyzing | - Detailing<br>Identifying the relevant elements of a subject matter from the irrelevant and essential parts from the unimportant. | Detailing biology as a science that studies life conceptually. |
| | - Structuring<br>Determining the systematic arrangement and structure between elements to form subject matter units. | Categorizing biology as a science that studies life conceptually. |
| Evaluating | - Assessing<br>Judging based on internal standards; the ability to access accuracy in reporting facts based on statements, documents, evidence, etc. | Conceptually assessing the nature of life |
| | - Criticizing<br>Judging based on external standards; the ability to compare work with the highest standards known in a field. | Critically the central theme in biology conceptually. |
| Creating | - Making<br>Finding a new product from the formulation and planning | Designing a concept map of hypothetical science processes metacognitively. |
| | - Detailing<br>Identifying the relevant elements of a subject matter from the irrelevant and vital parts from the unimportant. | Making a concept map of the relationships between the main themes in biology conceptually. |

Biology is part of STEM (Science, Technology, Engineering, and Mathematics) (Jackson, Koenig, & Smith, 2017; M. H. Wake, 2008). Biology should be empowered in the 21$^{st}$ century in order to integrate with the latest technology. Biological empowerment can be done through research, education and learning (Robinson et al., 2010). A general biology course is a subject that is closely related to life materials in the neighborhood. General biological material is very dense and wide so that in the delivery of many there are missed material because the delivery of material is too short. As a result, students are difficult to understand the material submitted both during the lecture and discussion students.

## METHOD

### General Background

This study used the development research designed to obtain a HOTS (transfer of knowledge) measurement instrument product. These development steps refer to the development stages including (1) designing test, (2) trial, (3) determining validity, and (4) determining reliability. First, the HOTS test instrument was designed and then validated by five experts, Dr. M. Masykuri, M.Si.; Dr. Baskoro Adi Prayitno, M.Pd.; Dr. Sumarwati, M.Pd.; Dr. Nonoh Siti Aminah, M.Pd.; Puguh Karyanto, M.Si., Ph.D. The content of all test items developed was validated by experts to meet the requirements of relevance. The questions were arranged in the form of descriptions for all indicators. Second, the items that had passed the content validation stage were tested on the participants or samples. The items that had been through content validation have been considered as meeting satisfactory quality content requirements. After going through the content validity stage, the instrument was then tested to ensure that each item in the instrument has validity and reliability according to the theory. Also, the things developed have proportional discriminating power and difficulty levels. In this second stage, the participants worked on a multiple-choice test of 12 items with four alternative answer choices. The trial results were analyzed and evaluated with the item discriminating power and difficulty level parameters. The analysis of the two parameters is based on the subject/participant response data to the test items.

The third and fourth steps as the final steps in the development of the instrument are to determine validity and reliability. From both tests, we can find out the magnitude after the data from the trial results are analyzed. The test result data is analyzed through the construct validity, the discriminating power of the item, and the difficulty level of the item to determine the validity and reliability of the instrument.

Validity is the essential characteristic in a measurement that refers to the accuracy of the test measurement function in question. Validating the test means to look for empirical evidence that the measurement results from the test do provide accurate information about the attributes being measured, without being tainted by irrelevant information (Cresswell, 2014). Reliability is the consistency of the measurement which means that the difference in scores obtained in the analysis indeed reflects the difference in the real ability, not the gap caused by the measurement error (Azwar, 2016; Cresswell, 2014).

### Sample/Participants/Group

The participants involved in this study consisted of five experts and 110 pre-service biology teachers. Some of the experts involved are the expert's biological instrument and material evaluation, who are registered as lecturers at state universities while the pre-service biology teachers are those listed as first-semester students in the 2018/2019 academic year under the faculty of teacher training and education — all the participants, including registered experts in state universities Central Java Province Indonesia.

### Data Analysis

The data were obtained from the test scores of the trial results. The number of HOTS items tested is 12, with a score of 1 (correct answer) and 0 (incorrect answer). Test

scores on the trial results were analyzed to determine the item's discriminating power, item difficulty level, construct validity, and item reliability. The instruments were arranged based on the outline of HOTS built based on the theory as in the following Table 2.

Table 2
HOTS Outline

| Aspect | Indicator | Item Number |
|---|---|---|
| Analyzing | Detailing biology as a science that studies life conceptually. | 1, 2 |
| | Categorizing biology as a science that studies life conceptually. | 3, 4 |
| Evaluating | Conceptually assessing the nature of life | 5, 6 |
| | Critically criticizing the central theme in biology conceptually. | 7, 8 |
| Creating | Designing a concept map of hypothetical science processes metacognitively. | 9, 10 |
| | Making a concept map of the relationships between the main themes in biology conceptually. | 11, 12 |

The item discrimination power can be done by calculating the correlation coefficient between the distribution of item scores and the total score. The correlation coefficient is a common variation between the scores of the two distributions that can be shown by the magnitude and direction of the correlation because the scores of the items in the test are dichotomous (there are only two kinds, namely 1 for the correct answer and 0 for the incorrect answer). Meanwhile, the distribution of the test scores is an interval number distribution. Then, the total item correlation coefficient can be calculated using the point-biserial correlation formula (r-pbis) (Azwar, 2016).

The magnitude of the correlation coefficient (r-pbis) moves from -1.00 to +1.00. In the item analysis, the discriminating power is usually considered good if the total item correlation coefficient reaches 0.30 or more, while the negative correlation coefficient is unacceptable (Azwar, 2016).

The item difficulty level is a parameter that describes how difficult it is for a group of subjects tested to give the correct answer to an item. This difficulty level of the item can be determined through the item difficulty index (p) (Azwar, 2016). In general, an index p around 0.50 is considered the best because p = 0.50 will produce the greatest item variance.

The two formulas above are for manually calculating item discriminating power and difficulty level. There is a computer program that can count both at once. One of the popular programs is ITEMAN 3.0. The magnitude of the correlation coefficient (r-pbis) can be determined through the output in the Item Statistics - Point Biser column while the item difficulty index (p) is shown in the Item Statistics - Prop column. Correct (Azwar, 2016).

The construct validity proves whether the measurement results obtained through the test items are highly correlated with the theoretical construct on which the test is constructed. Does the score obtain to support the theoretical concepts desired by the original measurement objectives (Azwar, 2016). The construct validity used in this study

is factorial validity with the confirmatory factor analysis (CFA) procedure. Through CFA, indicators that become the variables of the study can be directly measured. Because this study has three latent variables, which are also indicators of HOTS, the second-order CFA is used. To measure the value of each indicator, we can use the LISREL 8.50 computer program. Indicator criteria can be known from the t-value and standardized loading. The indicators are said to be valid if each indicator has t-value > 1.96 and a standardized loading value > 0.7 (Ghozali & Fuad, 2014) or standardized loading > 0.3 (Budiyono, 2019).

Through the LISREL 8.50 program, it can also be seen the reliability of an indicator by observing the squared multiple correlation (R2) value of the indicator (Ghozali & Fuad, 2014). The R2 value explains how much the proportion of indicator variance is explained by latent variables. The indicators are said to have reliability if the R2 value > 0.7. Moreover, the reliability coefficient is still allowed with a minimum amount of 0.65. The same thing can be done by ITEMAN 3.0 program, in which the reliability coefficient of the instrument can be seen from the Alpha output magnitude (Budiyono, 2019).

## FINDINGS

### Content Validity of the HOTS Instrument

By the method described above, content validity is the initial validity of the instrument development before proceeding to construct validity. The content validity was carried out by five experts who assessed the suitability of the items developed with the cognitive domains created. It aims to evaluate the construction of relevant items from the instrument of the study produced. The results of content validity by experts are displayed as the Aiken index of each item in Table 3 below.

Table 3
Aiken Index of HOTS Instrument

| Item | Rater | | | | | Score | Remark |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | Valid |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | Valid |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | Valid |
| 4 | 1 | 1 | 1 | 1 | 0 | 0.8 | Valid |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | Valid |
| 6 | 1 | 0 | 1 | 1 | 1 | 0.8 | Valid |
| 7 | 0 | 1 | 1 | 1 | 1 | 0.8 | Valid |
| 8 | 1 | 1 | 0 | 1 | 1 | 0.8 | Valid |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | Valid |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | Valid |
| 11 | 1 | 1 | 1 | 0 | 1 | 0.8 | Valid |
| 12 | 1 | 0 | 1 | 1 | 1 | 0.8 | Valid |

Remark:
Score 1: If the item fits into the cognitive domain of the construct.
Score 0: If the item does not fit into the cognitive domain of the construct.

Based on the results presented in Table 3, all items are valid. This is by the criteria stated by Guilford (1956) that if the Aiken index is less than 0.4, then the validity is low, and if more than 0.8, the validity is very high. The results of the content validity, in general, showed that all items were feasible for use after revising several things continued with trials. The inter-rater agreement (reliability) can be seen by calculating the inter-rater reliability coefficient using the Kappa coefficient (κ). The calculation results are presented in Table 4 below.

Table 4
Kappa (κ) Coefficient between Raters

| | | Rater | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Rater | 1 | | | | | |
| | 2 | 0.64 | | | | |
| | 3 | 0.75 | 0.64 | | | |
| | 4 | 0.75 | 0.64 | 0.75 | | |
| | 5 | 0.75 | 0.64 | 0.75 | 0.75 | |

Remark:
(κ) < 0.40: Not Good; 0.40 < (κ) < 0.70: Good; (κ) > 0.70: Very Good

Inter-rater reliability in Table 6 is an agreement between rater one and another. The mean of Kappa coefficient (κ), which is 0.70 is included in both categories. The reliability coefficient value of the HOTS test instrument obtained by the minimum criteria used is 0.70 (Miller, Linn, & Gronlund, 2009). The reliability values mean that the items in the instrument meet the reliability requirements.

**Construct Validity of HOTS Instrument**

The construct of HOTS theory was validated using the LISREL 8.50 application. Table 5 and Table 6 below are the results of the HOTS construct validation seen from the standardized solution and t-values file.

Table 5
Confirmatory Factor Analysis (CFA) Model from 12 indicators

| Aspect | Index | Description |
|---|---|---|
| Chi-Square | 55.06 | Fit Model |
| p-value | 0.225 | Fit Model |
| RMSEA | 0.037 | Fit Model |

Based on Table 5 above, the model fits with the category of P-Value = 0.22 (P > 0.05, which means not significant), RMSEA = 0.037 (Model fits if RMSEA <0.05) and Chi-value Square = 55.06. The CFA criterion means that the unidimensional aspect is met and can be continued to perform item analysis.

Table 6
Indicators and Standardized Loading Factors

| Factor | Indicator | Standardized Loading Factor Value | Value of $t_{obs}$ |
|---|---|---|---|
| Analyzing | A1 | 0.71 | ** |
| | A2 | 0.77 | 9.19 |
| | A3 | 0.75 | 8.92 |
| | A4 | 0.68 | 8.17 |
| Evaluating | E1 | 0.81 | ** |
| | E2 | 0.68 | 7.60 |
| | E3 | 0.74 | 8.70 |
| | E4 | 0.71 | 7.33 |
| Creative | K1 | 0.81 | ** |
| | K2 | 0.78 | 12.62 |
| | K3 | 0.67 | 7.69 |
| | K4 | 0.65 | 8.08 |

By looking at the values in Table 6, all indicators have a standard loading of more than 0.5 and all values of the observations more than 1.96. According to this data, it can be concluded that all indicators have functioned well in supporting the construct validity (Budiyono, 2019). It also means that the relationship between the hypothesized variables is supported by empirical data, and all items meet the element of validity.

The test results prove that the six indicators developed are valid for the measurement of HOTS constructs because the model is by empirical data based on the CFA measurement model criteria. Thus, this measurement instrument can be said to have fulfilled the unidimensional assumption so that it can proceed to the analysis stage of the item discriminating power and difficulty level.

**Item Discriminating Power**

The results of the item analysis based on the discriminating power of the items showed that the average discrimination index of all items is 0.54. This means that the items developed have a good discriminating power category that can distinguish participants with high and low cognitive abilities. Table 7 below describes the item discrimination power index of 12 questions.

Table 7
Item Discriminating Power Index

| Item | Point-biserial correlation (r-pbis) | Remark |
|---|---|---|
| 1 | 0.58 | Good |
| 2 | 0.68 | Good |
| 3 | 0.66 | Good |
| 4 | 0.50 | Good |
| 5 | 0.44 | Good |
| 6 | 0.56 | Good |
| 7 | 0.39 | Good |
| 8 | 0.53 | Good |
| 9 | 0.66 | Good |
| 10 | 0.54 | Good |
| 11 | 0.54 | Good |
| 12 | 0.43 | Good |
| Mean | 0.54 | |
| SD | 0.09 | |

Based on Table 7 above, all items have a good category of discriminating power index with a mean of 0.54 and a standard deviation of 0.09. The highest discriminating power index is in item number 2 of 0.68 while the lowest in item number 7 of 0.39.

**Item Difficulty Level**

In addition to the item discriminating power index, the HOTS instrument developed also analyzed the difficulty level of the item. The following Table 8 shows the difficulty level of all items.

Table 8
Item Difficulty Level

| Item | Item Difficulty Index (p) | Category |
|---|---|---|
| 1 | 0.51 | Medium |
| 2 | 0.32 | Medium |
| 3 | 0.33 | Medium |
| 4 | 0.52 | Medium |
| 5 | 0.49 | Medium |
| 6 | 0.51 | Medium |
| 7 | 0.50 | Medium |
| 8 | 0.46 | Medium |
| 9 | 0.26 | Medium |
| 10 | 0.34 | Medium |
| 11 | 0.29 | Medium |
| 12 | 0.45 | Medium |
| Mean | 0.42 | |
| SD | 0.09 | |

By Table 8 above, it is evident that all items have a medium difficulty level with a mean of 0.42 and a standard deviation of 0.09. The item with the highest difficulty level is item number 4 (0.52), while the lowest is number 9 (0.26).

## DISCUSSION

Higher-order thinking skills (transfer of knowledge) are a hot topic in the 21st-century era. Therefore, we need the right instrument to measure these skills. Based on the results presented from the initial stages of the design to construct validation, several findings require further discussion.

The initial stage of this study is to design the instrument to adjust to the construct of the theory developed. HOTS as a transfer of knowledge refers to the high-order cognitive domain of the revised Bloom's Taxonomy, namely C4 analysis, C5 evaluation, and C6 creation. However, the researcher formulated his theory by simplifying the indicators from each aspect of the cognitive domains. The symbols of each area are as follows: analyzing has the signs of detailing and structuring; evaluating has the symbols of assessing and criticizing, and creating has the indicators of making and describing. The construct of the development of such theory needs to be validated to ensure that each indicator is a component of the variables developed.

After going through the design stage, the HOTS (transfer of knowledge) instrument was validated by the experts. The final results of the validation of the experts showed that the assessment instrument had met the valid categories and was ready to be used in the data collection trial. The assessment tool is based on reliable and relevant supporting theories. The design of the HOTS instrument declared valid and appropriate to be used to test the construct validity of the scientific reasoning test is determined with the confirmatory factor analysis (CFA) method. The use of CFA is to check dimensionality as a reference to verify the unidimensional assumptions of the measurement instrument. At this stage, the analysis was carried out using the Lisrel version 8.5 program. Analysis of the construct validity with CFA was carried out with the second-order confirmation analysis. In this study, HOTS consist of three aspects, namely analysis, evaluation, and creation, so it is necessary to test the suitability of HOTS models that are by empirical data using CFA.

The test results on the CFA second-order measurement model with 12 items produce P-Value = 0.22503 ($P > 0.05$) and RMSEA = 0.037 ($RMSEA < 0.05$). Based on the data, it can be concluded that the model fits with empirical data. In other words, the instrument that measures HOTS is designed to meet the construct validity. Also, all indicators have a standard loading of more than 0.5, and all observed values more than 1.96. This proves that each item developed is valid for HOTS measurement.

A good instrument is the one that meets the aspects of validity and reliability. All items that have been developed have met the validity aspect based on the above criteria. To find out the reliability aspect, we can see the magnitude of the reliability coefficient at the ITEMAN 3.0 program output with the Alpha technique (Budiyono, 2018). The importance of Alpha is calculated at 0.784. This proves that the instrument developed has a reliable category. The reliability coefficient is in the range of 0 and 1. If the reliability coefficient gets greater (close to the value of 1), then the error variance gets smaller, and the resulting score of the item is closer to the actual rating (Budiyono, 2019). Thus, it can be concluded that the HOTS instrument is valid and reliable.

Because this measurement instrument has fulfilled unidimensional assumptions based on construct validation, the analysis of discriminating power and the difficulty of the item can be made. An item has an excellent discriminating power if the smart group answers the item more correctly than the not talented group. The items are said to have an extraordinary discriminating power if the selective power index is equal to or more than 0.30 (Budiyono, 2018). Based on the results of the analysis of ITEMAN 3.0, all items have a suitable category of discriminating power index with a mean of 0.54 and a standard deviation of 0.09. The highest discriminating power index is in item number 2 of 0.68 while the lowest in item number 7 of 0.39. It can be concluded that the HOTS instrument developed has good discriminating power.

In addition to analyzing the item discriminating power, the difficulty level of the item is also investigated. This was done to maintain the quality of the developed HOTS instrument. The item difficulty level is a proportion of the number of participants who correctly answered the question items of all test participants (Budiyono, 2018). The difficulty index (P) allowed depends on the urgency and purpose of the instrument being developed. Given this instrument for measuring HOTS, the interval used is $0.25 \leq P \leq 0.75$ (Budiyono, 2018). Based on the results of the analysis of the item difficulty level, all items have a medium category of difficulty level with a mean of 0.42 and a standard deviation of 0.09. The question with the highest difficulty level is item number 4 (0.52), while the lowest is number 9 (0.26). All items do not have a difficulty index close to 0, which means too tricky or close to 1, which means too easy.

## CONCLUSIONS AND IMPLICATIONS

Based on the results of data analysis, it can be concluded that the HOTS instrument is ready to be used as a measurement instrument of HOTS (transfer of knowledge) in the biology lecture. The data of the validity test results by the validator on each item of the measurement instrument indicate that all instrument items are eligible to be used as measurement instruments. The results of construct validation, item discriminating power analysis, and item difficulty level showed all elements of the HOTS instrument in biology lecture are of proper quality criteria.

The results of this study can have implications for curriculum development in biology lectures and are used as an assessment model that can facilitate lecturers as a basis for improving expected learning outcomes. Therefore, lecturers must have the skills to develop assessment instruments, and students or pre-service teachers must be trained to have higher-order thinking skills in the learning process. Moreover, HOTS is part of $21^{st}$-century skills as a basis for dealing with the demands of life in the future.

Based on the findings of the characteristics of the instruments developed, it is recommended that lecturers be able to train students through lectures involving technology to develop higher-order thinking skills based on analytic, evaluative and creative domains. Lecturers must get used early on to apply HOTS tests in biology lectures that are still not optimal, so the HOTS of students as pre-service biology teachers continue to develop in another course.

**ACKNOWLEDGMENTS**

The research funding supported by LPDP (Lembaga Pengelola Dana Pendidikan) provided by the Ministry of Finance Indonesia. Any research progress should be reported through the monitoring and evaluation system.

**REFERENCES**

Afandi, & Sajidan. (2017). *Stimulasi keterampilan berpikir tingkat tinggi-konsep dan implementasinya dalam pembelajaran abad 21*. (Gunarhadi & Sumarwati, Eds.) (Edisi 1). Surakarta: UNS Press.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. United States of America: Addison Wesley Longman.

Azwar, S. (2016). *Konstruksi tes kemampuan kognitif* (1st ed.). Yogyakarta: Pustaka Pelajar.

Bilyalova, A. A., Salimova, D. A., & Zelenina, T. I. (2020). Digital transformation in education. In *ICIS 2019 Vol.1*, 265–276. Springer International Publishing.

Budiyono. (2018). *Pengantar penilaian hasil belajar*. (Suyono, Ed.) (1st ed.). Surakarta: UNS Press.

Budiyono. (2019). *Pengantar teori pengukuran pendidikan*. (M. Kurnianingtias, Ed.) (1st ed.). Surakarta: UNS Press.

Cresswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approach*. Los Angeles: Sage Publications Ltd.

Fryer, L. K., & Bovee, H. N. (2018). Staying motivated to e-learn: Person- and variable-centered perspectives on longitudinal risks and support. *Computers & Education*, *120*, 227-240. https://doi.org/10.1016/j.compedu.2018.01.006.

García-Martín, J., & García-Sánchez, J. N. (2017). Pre-service teachers' perceptions of the competence dimensions of digital literacy and psychological and educational measures. *Computers and Education*, *107*, 54–67.

Ghozali, I., & Fuad. (2014). *Structural equation modelling: teori, konsep dan aplikasi dengan program Lisrel 9.10*. Semarang: Badan Penerbit Universitas Diponegoro.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education*. New York: Mc Graw-Hill Book Co. Inc.

Heong, Y. M., Othman, W. B., Yunos, J. Bin, Kiong, T. T., Hassan, R. Bin, Mohaffyza, M., & Mohamad, B. (2011). The level of Marzano higher-order thinking skills among technical education students. *International Journal of Social Science and Humanity*, *1*(2), 121–125.

Jackson, H. E., Koenig, K., & Smith, L. M. (2017). *Enhancing engineering student learning in foundational STEM courses of biology, chemistry, mathematics, and physics: Transforming the faculty culture mathematics by transforming the faculty*

*culture*. American Society for Engineering Education.

Jeffrey, L. M., Milne, J., Suddaby, G., & Higgins, A. (2014). Blended learning : How teachers balance the blend of online and classroom components. *Journal of Information Technology Education*, *13*, 121–140.

Jeong, J., Kim, H., Chae, D., & Kim, E. (2014). The effect of a case-based reasoning instructional model on Korean high school students' awareness in the climate change unit. *Eurasia Journal of Mathematics, Science & Technology Education*, *10*(5), 427–435.

Kong, S. C., Chan, T.-W., Griffin, P., Hoppe, U., Huang, R., Kinshuk, Yu, S. (2014). E-learning in school education in the coming 10 years for developing 21st-century skills: Critical research issues and policy implications. *Journal of Educational Technology & Society*, *17*(1), 70–78.

Lee, J., Lim, C., & Kim, H. (2017). Development of an instructional design model for flipped learning in higher education. *Educational Technology Research and Development*, *65*(2), 427–453.

Liu, C.-Y., Wu, C.-J., Wong, W.-K., Lien, Y.-W., & Chao, T.-K. (2017). Scientific modeling with mobile devices in high school physics labs. *Computers & Education*, *105*, 44–56.

Masigno, R. M. (2014). Enhancing higher-order thinking skills in a marine biology class through problem-based learning. *Asia Pacific Journal of Multidisciplinary Research*, *2*(5), 1–6.

Mazalah, A., Jamaludin, B., Ahmad Zamri, M., Aidah, A. K., Fariza, K., Mohd Yusof, D., & Diana Fazleen, Z. (2016). The application of 21st-century ICT literacy model among teacher trainees. *Turkish Online Journal of Educational Technology*, *15*(3), 151–161.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching*. United States of America: Pearson Education, Inc.

Moore, J. C., & Rubbo, L. J. (2012). Scientific reasoning abilities of nonscience majors in physics-based courses. *Physical Review Special Topics-Physics Education Research*, *8*(010106), 1–8.

Nguyễn, T. M. T., & Nguyễn, T. T. L. (2017). Influence of explicit higher-order thinking skills instruction on students' learning of linguistics. *Thinking Skills and Creativity*, *26*, 113–127.

Oey-Gardiner, M., Rahayu, S. I., Abdullah, M. A., Effendi, S., Darma, Y., Dartanto, T., & Aruan, C. D. (2017). *Era disrupsi peluang dan tantangan pendidikan tinggi Indonesia*. (D. Dhakidae, Ed.). Jakarta Pusat: Akademi Ilmu Pengetahuan Indonesia.

P21 (Partnership For 21st Century Skills). (2012). *Learning for the 21st-century. A report and mile guide for 21st-century skills*. Retrieved from http://www.p21.org

Piaget, J., & Kohler, R. (2014). *Bloomsbury library of educational thought*. (R. Kohler,

Ed.). London, UK: Continuum International Publishing Group.

Prayitno, B. A., Suciati, & Titikusumawati, E. (2018). Enhancing students' higher-order thinking skills in science through instad strategy. *Journal of Baltic Science Education*, *17*(6), 1046–1055.

Robinson, G. E., Banks, J. A., Padilla, D. K., Burggren, W. W., Cohen, C. S., Delwiche, C. F., … Tomanek, L. (2010). Empowering 21st century biology. *BioScience*, *60*(11), 923–930.

Saido, G. A., Siraj, S., & Nordin, A. B. (2015). Teaching strategies scale for promoting higher-order thinking skills among students in science. *Proceedings of ISER 5th, International Conference*, (September), 91–94.

Shuler, C., Winters, N., & West, M. (2013). The future of mobile learning - implications for policymakers and planners. *UNESCO - United National Educational.*, 1–44.

Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275.

Tan, S. Y., & Halili, S. H. (2015). Effective teaching of higher-order thinking (HOT) in education. *The Online Journal of Distance Education and E-Learning*, *3*(2), 41–47.

Uerz, D., Volman, M., & Kral, M. (2018). Teacher educators' competences in fostering student teachers' proficiency in teaching and learning with technology: An overview of relevant research literature. *Teaching and Teacher Education*, *70*, 12–23.

Utama, C., Sajidan, Nurkamto, J., & Wiranto. (2019). Using TPACK as a framework to analyze TLC model. In *Journal of Physics: Conference Series* (Vol. 1175). Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/1175/1/012146.

van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior*, *72*, 577–588.

Wake, J. D., Guribye, F., & Wasson, B. (2018). Learning through collaborative design of location-based games. *International Journal of Computer-Supported Collaborative Learning*, (1), 1–21.

Yanto, B. E., Subali, B., & Suyanto, S. (2019). Measurement instrument of scientific reasoning test for biology education students. *International Journal of Instruction*, *12*(1), 1383–1398.

Yee, M. H., Lai, C. S., Tee, T. K., & Mohamad, M. M. (2016). The role of higher-order thinking skills in green skill development. *EDP Sciences*, *70*(05001), 1–5.

Zohar, A., & Cohen, A. (2016). Large scale implementation of higher-order thinking (HOT) in civic education: The interplay of policy, politics, pedagogical leadership, and detailed pedagogical planning. *Thinking Skills and Creativity*, *21*, 85–96.