**International Journal of Evaluation and Research in Education (IJERE)**
Vol. 10, No. 2, June 2021, pp. 545~552
ISSN: 2252-8822, DOI: 10.11591/ijere.v10i2.20796 

# Development of higher order thinking skill assessment instruments in learning Indonesian history

**Johan Setiawan, Ajat Sudrajat, Aman, Dyah Kumalasari**
Doctoral Program, Postgraduate Program, Yogyakarta State University, Indonesia

**Article Info**

***Article history:***

Received Oct 12, 2020
Revised Mar 26, 2021
Accepted Apr 17, 2021

***Keywords:***

Assessment instruments
Higher order thinking skill
Indonesian history learning

**ABSTRACT**

This study aimed to: 1) Produce higher order thinking skill (HOTS) assessment instruments in learning Indonesian history; 2) Know the validity of HOTS assessment instruments in learning Indonesian history; 3) Find out the characteristics of HOTS questions in learning Indonesian history. This study employed the research and development method of the Borg and Gall model. The HOTS test item was conducted on 36 students in class XI of 2 Ngaglik State Senior High School. Data analysis includes tests of validity, reliability, level of difficulty, distinguishing features and deception index. The study found: 1) The HOTS assessment instrument of multiple-choice questions consisted of 25 items; 2) The results of the HOTS question validation by two Indonesian history learning assessment experts on the material, construction and language aspects were valid and appropriate. The results of the validation by three Indonesian history teachers also stated that the assessment instruments were valid and appropriate; 3) The characteristics of HOTS questions had fulfilled the validity criteria of 23 questions, reliability with a coefficient of 0.97 (very strong), the average difficulty level is 0.33 (moderate), the average differentiation test is 0.42 (good), and the average deception index is 0.56 (good).



***Corresponding Author:***

Johan Setiawan
Doctoral Program, Postgraduate Program
Yogyakarta State University
Yogyakarta, Indonesia
Email: johansetiawan.2019@student.uny.ac.id; johansetiawan767@gmail.com

## 1. INTRODUCTION

Learning in the 21st century now emphasizes the ability to think critically and solve problems [1]-[3]. This becomes essential that must be mastered by students, of course, requires the concept of assessment that is able to describe the desired learning objectives. The assessment standard in the 2013 curriculum is done by adapting international standardized assessment models, one of the international standardized assessment models is higher order thinking skills (HOTS) [4], [5]. HOTS functions to assess whether students already have high-level thinking skills such as: C4 (analyzing), C5 (evaluating), and C6 (making) [6]-[9].

HOTS is a quality thinking ability that is conceptually based on Bloom's Taxonomy level of thinking [10], he argues that students not only need remembering skills, but must have higher thinking abilities to overcome increasingly complex problems and have critical and rational thinking abilities [11], [12]. Learning done in schools should not only remember the concepts and knowledge, but analyze, evaluate and create the problems faced [13], [14]. Students must often be faced with exercises working on HOTS questions that are interesting to solve, so that the potential of students increases.

In the current 2013 curriculum, while the syllabus serves as a guide in learning activities in schools, the material prepared requires students to have high-level thinking skills [15], and of course the assessment must adjust to the ability to think at a high level, especially in history learning in high school [16], [17]. The meaning of assessment in the realm of education is to find out the achievement of objectives and the process of learning activities [18], [19], therefore assessment must meet academic requirements as an appropriate assessment carried out in high school.

Assessments used by Indonesian history teachers often do not help students optimally in dealing with contextual problems [20]-[22]. There are still many teachers who make test questions not based on the test grid, but tend to only use questions on books that are provided [23]. The tendency of teachers to make questions that are not guided by the test grid is what makes students not trained in high-level thinking [24]. High-level thinking skills of students in Indonesia are very low, especially in history learning. This can be seen from the ability of students to conduct investigations, understand theory, analysis, and solve problem [25], [26].

The use of HOTS assessment instruments in learning Indonesian history, because in the 2013 curriculum only has an allocation of two hours per week, the target to be achieved requires an effective and efficient strategy to meet the learning objectives. These targets must meet the appropriate learning base and good instruments as an assessment of their achievements. Learning assessment in the 2013 curriculum includes: 1) Knowledge with written tests, observations and assignments; 2) Skills with performance, projects, products, portfolios; and 3) Attitudes by observation, self-assessment, and journals [27]. Various aspects of the assessment, Indonesian history teachers must be creative in getting around the allocation of hours distribution so that competence is achieved [28].

Based on the results of interviews with Indonesian history teachers at 2 Ngaglik State Senior High School, that the instruments used still measure aspects of memorization and understanding. Learning Indonesian history requires assessment instruments that can train high-level skills, such as material in basic competence 3.6 analyze the role of national and regional figures in fighting for Indonesian independence. This material has broad discussion and requires a lot of student activities in the classroom including high-level thinking activities, this is the basis for researchers to develop HOTS assessment instruments [29].

Previous research on higher order thinking skills has been carried out by several researchers who have focused on learning mathematics [30], learning physics [31] and learning history with descriptive questions [17]. Based on previous research studies on the HOTS, researchers found a new aspect of this research is the development of HOTS assessments with multiple choice test questions, especially in assessing the learning of Indonesian history. Based on the background that has been described, the researcher is interested in conducting research under the title development of higher order thinking skill (HOTS) assessment instruments in learning Indonesian history in high schools.

## 2. RESEARCH METHOD

### 2.1. Type of research

This type of research is a Research and Development (R&D) using the Bord and Gall development model [32], which has been modified into five stages of research in accordance with the objectives and interests in this study. The stages consist of: 1) Needs analysis and preliminary information gathering; 2) Planning and preparation of assessment instruments; 3) Initial product testing by experts; 4) Evaluation; and 5) Implementation.

### 2.2. Research design

To examine the appropriateness of HOTS assessment instruments in learning Indonesian history on the material of resistance of the Indonesian people to European colonization until the 20th century, it was first validated by instrument experts and evaluation experts [33], [34], then revised in stage one. The revised product was then validated by three Indonesian history teachers, then a second stage revision was carried out. The second stage revised product was tested on one class at 2 Ngaglik State Senior High School. This school is located in Sleman District, Special Region of Yogyakarta, Indonesia.

### 2.3. Research subjects

Research subjects on higher order thinking skills assessment instrument products in the learning Indonesian history on the material of resistance of the Indonesian people against European colonization until the 20th century were conducted on population of class XI IPS 3 at 2 Ngaglik State Senior High School as many as 36 students. The research was conducted on December 12 until January 31, 2019.

### 2.4. Data collection techniques and instruments

Data collection techniques used questionnaires and tests [35], questionnaires were used to measure responses of variables to experts in the form of high-level thinking skills (HOTS) in learning Indonesian history on material of Indonesian resistance to colonialism in Europe until the 20th century. Data collection instruments consist of: 1) Test instruments, tests in the form of multiple choice (dichotomy 0 and 1) with five answer choices [36], [37], which refers to indicators of high-level thinking ability totalling 25 questions; 2) Validation sheets, carried out by two experts namely: instrument experts for the validation of test instruments and evaluation experts for validation of HOTS assessments [38]-[40]. Furthermore, validation was carried out by three Indonesian history teachers to determine the practicality and response of teachers as instructors in the school environment.

### 2.5. Data analysis techniques

Data analysis was performed to obtain valid and reliable HOTS assessment instruments [41]. Data analysis is done in two ways, namely:

#### 2.5.1. Qualitative data analysis of the results of validation sheets

Qualitative analysis of HOTS test questions was obtained from the results of logical validation sheets based on three aspects namely: material, construction and language aspects [42]-[45]. Valid test items are used based on the validator assessment of two expert lecturers and three Indonesian history teachers. The values given to each item of validation are: value one "invalid", value two "less valid", value three "quite valid", value four "valid" and value five "very valid". Analysis of HOTS test items was calculated using the Aiken's V calculation formula.

#### 2.5.2. Quantitative data analysis of HOTS test questions

Data obtained from students' responses were analyzed using the Microsoft Excel program. The analysis was carried out to determine the characteristics of HOTS items which included:

A. Validity test
   The validity test of the test using the biserial point correlation formula with the provisions of the calculation results compared with $r_{table}$ at 5% significance level. If $r_{count}$ is greater or equal to $r_{table}$ then the item is valid, but if $r_{count}$ is smaller than $r_{table}$ then the item is invalid [46].
B. Reliability test
   The reliability test uses the KR-20 formula, because the score questions are dichotomous (0 and 1). To determine the reliability criteria if the interval value of the coefficient ≥0.7 [47], [48].
C. Test difficulty level
   Difficulty level analysis of test questions is needed to examine the items in terms of difficulty, so that items can be obtained that fall into the category of easy, medium and difficult. The formula used to calculate the difficulty of items is: $P=N_P/N$ [49] by using the criteria: 0.00-0.30 "too difficult", 0.31-0.70 "moderate", and 0.71- 1.00 "too easy".
D. Distinguishing power
   Distinguishing power is the ability of the test to distinguish between students who have high abilities and students with low ability. Different item power index using the formula DP=BA/JA-BB/JB [50]. If the distinguishing index is known, then the number is interpreted on the criteria: 0.00-0.20 "bad", 0.21-0.40 "sufficient", 0.41-0.70 "good" and 0.71-1.00 "very good".
E. Deception index
   In the case of multiple-choice forms there are alternative answers which are deceitful, deceitful here is the answer to the question that can deceive the real answer. Using the formula IP=P x 100/(N-B)/(n-1) [51] with the criteria: 76%-124% "very good", 51%-75% or 126%-150% "good", 26%-50% or 151%-175% "not good", 0%-25% or 176%-200% "bad", and >200% "misleading".

## 3. RESULTS

### 3.1. Data of product test result

#### 3.1.1. Expert validation

Validation is done by providing a text in the form of a validation sheet to the instrument experts and evaluation experts, then analyzed using the Aiken's V formula to calculate the content validity coefficient. The results of the validation analysis of the instrument experts and evaluation experts are shown in Table 1. Table 2 shows that the calculation of Aiken's V coefficient based on the validation of the instrument expert consisting of 10 questionnaire items and evaluation experts consisting of 20 questionnaire items, all declared to be eligible to use.

Table 1. Results of instrument expert validation analysis

| Results of instrument expert analysis | | | | | |
|---|---|---|---|---|---|
| Question number | Aiken's V coefficient | Criteria | Question number | Aiken's V coefficient | Criteria |
| 1 | 1.00 | Eligible to use | 6 | 0.50 | Eligible to use |
| 2 | 0.75 | Eligible to use | 7 | 0.75 | Eligible to use |
| 3 | 0.75 | Eligible to use | 8 | 1.00 | Eligible to use |
| 4 | 1.00 | Eligible to use | 9 | 1.00 | Eligible to use |
| 5 | 0.75 | Eligible to use | 10 | 1.00 | Eligible to use |

Table 2. Result of validation of expert evaluations

| Result of expert evaluation analysis | | | | | |
|---|---|---|---|---|---|
| Question number | Aiken's V coefficient | Criteria | Question number | Aiken's V coefficient | Criteria |
| 1 | 1.00 | Eligible to use | 11 | 0.75 | Eligible to use |
| 2 | 1.00 | Eligible to use | 12 | 1.00 | Eligible to use |
| 3 | 1.00 | Eligible to use | 13 | 1.00 | Eligible to use |
| 4 | 0.75 | Eligible to use | 14 | 1.00 | Eligible to use |
| 5 | 0.75 | Eligible to use | 15 | 1.00 | Eligible to use |
| 6 | 1.00 | Eligible to use | 16 | 1.00 | Eligible to use |
| 7 | 0.75 | Eligible to use | 17 | 1.00 | Eligible to use |
| 8 | 1.00 | Eligible to use | 18 | 0.75 | Eligible to use |
| 9 | 1.00 | Eligible to use | 19 | 1.00 | Eligible to use |
| 10 | 1.00 | Eligible to use | 20 | 1.00 | Eligible to use |

### 3.1.2. Validation by Indonesian history teacher

Validation was carried out to see the contents and effectiveness of the initial product to three Indonesian history teachers, then analyzed the HOTS test items according to the validator's assessment using Aiken's V formula to calculate the content validity coefficient. The validation analysis data are as shown in Table 3.

Table 3. Results of validation analysis of Indonesian history teacher

| Question | Rater 1 | Rater 1 | Rater 1 | Aiken's V Coefficient | Criteria |
|---|---|---|---|---|---|
| 1 | 5 | 5 | 4 | 0.91 | Eligible to use |
| 2 | 5 | 5 | 5 | 1.00 | Eligible to use |
| 3 | 4 | 5 | 4 | 0.83 | Eligible to use |
| 4 | 4 | 4 | 4 | 0.75 | Eligible to use |
| 5 | 5 | 4 | 4 | 0.83 | Eligible to use |
| 6 | 4 | 4 | 5 | 0.83 | Eligible to use |
| 7 | 5 | 3 | 5 | 0.83 | Eligible to use |
| 8 | 5 | 3 | 5 | 0.91 | Eligible to use |
| 9 | 5 | 4 | 5 | 0.66 | Eligible to use |
| 10 | 4 | 3 | 4 | 0.75 | Eligible to use |
| 11 | 3 | 4 | 4 | 1.00 | Eligible to use |
| 12 | 5 | 4 | 5 | 0.91 | Eligible to use |
| 13 | 5 | 4 | 4 | 0.83 | Eligible to use |
| 14 | 5 | 3 | 4 | 0.66 | Eligible to use |
| 15 | 4 | 4 | 4 | 0.83 | Eligible to use |
| 16 | 4 | 3 | 4 | 0.83 | Eligible to use |
| 17 | 5 | 3 | 5 | 1.00 | Eligible to use |
| 18 | 4 | 4 | 5 | 0.83 | Eligible to use |
| 19 | 5 | 4 | 4 | 0.83 | Eligible to use |
| 20 | 4 | 4 | 5 | 0.83 | Eligible to use |

Based on Table 3, the results are obtained that all multiple-choice items consisting of 30 HOTS items are in the valid category with the lowest index 0.66 and the highest 1.00. The interpretation is done by using criteria less than 0.6 then the validity is said to be low, between 0.6-0.8 in the moderate category and if more than 0.8 is said to be high.

### 3.2. Limited trial result data

A limited trial was conducted at 2 Ngaglik State Senior High School, Yogyakarta by involving students of class XI IPS 3 totalling 36 students. The quality of HOTS test questions based on the characteristics of the questions include: validity, reliability, level of difficulty, distinguishing features and deception index. The results of the interpretation of the item analysis are:

### 3.2.1. Item validity

Validation of the test consists of 30 multiple choice items which are calculated using the Microsoft Excel program, then interpreted with $r_{table}$ at a significance level of 5% and N=36. Then obtained $r_{table}$ of 0.339 and there are 23 valid questions. The results of the validation of the questions are as shown in Table 4.

Table 4. Validity test results

| Question | Validity index | Question item | Total | Percentage |
|---|---|---|---|---|
| 1 | >0.339 (Valid) | 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 | 23 | 92% |
| 2 | ≤0.339 (Invalid) | 9, 14 | 2 | 8% |

Based on Table 4, it can be concluded that the HOTS question in the study of Indonesian history on the material of Indonesian people resistance against European colonization until the 20th century has good quality in terms of its validity because the number of valid items is more than 50%. This means that the HOTS question can measure what should be measured.

### 3.2.2. Item reliability

Testing the reliability of HOTS questions using the KR-20 formula with the help of the Microsoft Excel program, the calculation results show the reliability of the questions of 0.97 so that the items have a very strong level of reliability in the category. The item reliability results are as shown in Table 5.

Table 5. Reliability test results

| Reliability score | Criteria |
|---|---|
| 0.97 | Very strong |

### 3.2.3. Item difficulty

The difficulty of HOTS items is calculated using the formula $P=N_P/N$ with the help of the Microsoft Excel program, the calculation results show seven questions are classified as too difficult, 16 questions are classified as moderate and two questions are classified as too easy. The results of the difficulty of the items are as shown in Table 6.

Table 6. Results of item difficulty

| Question | Difficulty index | Item | Total | Percentage |
|---|---|---|---|---|
| 1 | 0.00-0.30 | 4,5,10,12,15,17,18,21,23 | 9 | 36% |
| 2 | 0.31-0.70 | 1,2,3,6,7,8,9,11,13,14,16,19,20,22 | 14 | 56% |
| 3 | 0.70-100 | 24,25 | 2 | 8% |

### 3.2.4. Item distinguishing power

The distinguishing power test about HOTS is calculated using Microsoft Excel program. The calculation results are as shown in Table 7.

Table 7. Distinguishing power results

| Question | Distinguishing power | Item number | Total | Percentage |
|---|---|---|---|---|
| 1 | 0.00-0.20 | - | - | 0% |
| 2 | 0.21-0.40 | 1,4,7,8,15,22 | 6 | 24% |
| 3 | 0.41-0.70 | 2,3,5,6,9,10,12,16,19,21,24 | 11 | 44% |
| 4 | 0.71-1.00 | 11,13,14,17,18,20,23,25 | 8 | 32% |

### 3.2.5. Index of deception item

Deception index analysis about HOTS is calculated employing the Microsoft Excel program. The results of the deception index as shown in Table 8.

Table 8. Results of the deception index

| Question | Answer Option | Number of options selected | Percentage | Criteria |
|---|---|---|---|---|
| 1 | A, B, C*, D, E | A=3, B=2, C=18, D=4, E=3 | A=62.5%, B=76%, C=60%, D=83%, E=62.5% | Good |
| 6 | A, B, C, D, E* | A=2, B=4, C=3, D=4, E=17 | A=57%, B=70%, C=57%, 82%, D=70%, E=57% | Good |
| 14 | A*, B, C, D, E | A=15, B=4, C=3, D=5, E=3 | A=52%, B=66%, C=50%, D=83%, E=50% | Good |

## 4. DISCUSSION

This research resulted in the development of HOTS assessment instruments that can be used in learning Indonesian history in high schools. The HOTS assessment instrument in the study of Indonesian history focuses on one of the materials, namely the material of Indonesian people resistance against European colonization until the 20th century. HOTS questions in this study consisted of 25 multiple choice questions with five answer choices that included levels of C4 (analyzing), C5 (evaluating), and C6 (making). The HOTS concept used in this study refers to the revised high-level thinking concept of Bloom's Taxonomy.

The validation of HOTS assessment instruments using logical validation that includes aspects of material, construction and language analyzed in accordance with the validator's assessment using the Aiken's V formula to calculate content validity coefficient. The results of two Indonesian history learning assessment experts show HOTS assessment instruments are valid and appropriate to use. The results of the validation by three Indonesian history teachers also stated that the HOTS assessment instrument was valid and proper to use.

The characteristics of multiple-choice items calculated using the help of the Microsoft Excel program show that of the 25 items, there are two invalid questions. These results indicate that the quality of validity is good, because the number of valid items is more than 50%. This means that the HOTS questions can measure what should be measured. The average level of item reliability is 0.97, so it can be concluded that the item has a very strong level of reliability. The level of difficulty of the items was an average of 0.33 in the medium category. The average of distinguishing power test is 0.42 in the good category, and the average of deception index is 0.56 in the good category.

This assessment instrument can be used to measure high-level thinking skills in high school students in learning Indonesian history. Especially in the material of Indonesian people resistance against the European occupation until the 20th century. Tests containing high-level thinking questions (HOTS) encourage students to think about subject matter [52].

## 5. CONCLUSION

The conclusions obtain: 1) HOTS assessment instruments in learning Indonesian history on the material of Indonesian people resistance against European colonization until the 20th century, consisting of 25 multiple choice questions with five answer choices; 2) Validation of HOTS questions are shown from the results of the analysis validator conducted by two assessment experts and three Indonesian history teachers. The results of the expert assessment analysis show the HOTS assessment instrument is feasible to use, and also the results of the validation analysis of three Indonesian history teachers also show that the HOTS assessment instrument is feasible to use; 3) The characteristics of multiple choice items show that of the 25 items, there are two questions that are not valid, the average of reliability level is 0.97 (very strong category), the average of difficulty level is 0.33 (medium category), the average of distinguishing power test is 0.42 (good category), and the average of deception question index is 0.56 (good category).

This HOTS assessment instrument product of multiple-choice questions has met the eligibility standards of items in terms of validity, reliability, difficulty level, distinguishing features and deception index. The implication is that these HOTS questions can be used by students as training material to practice high-level thinking skills and as an alternative assessment instrument to assist teachers in preparing HOTS-based questions that will be implemented to students. The limitations of this research are: Product development trials have not been carried out at the field trial stage, this research is still at the limited trial stage involving one class at SMAN 2 Ngaglik. Recommendations for further research should be carried out field trials involving a wider range of respondents. The development of HOTS assessment instruments in the form of multiple choices is still limited to one of the historical KD (Basic Competencies) of class XI. It is necessary that further research can develop HOTS assessment instruments on other basic competencies.

## REFERENCES

[1] A. Chalkiadaki, "A systematic literature review of 21st century skills and competencies in primary education," *International Journal of Instruction*, vol. 11, no. 3, pp. 1-16, 2018.

[2] W. Conklin, *Higher order thinking skills to develop 21 century learners*. California: Shell Education, 2012.

[3] M. C. Quieng, P. P. Lim, and M. R. D. Lucas, "21st century-based soft skills: Spotlight on non-cognitive skills in a cognitive-laden dentistry program," *European Journal of Contemporary Education*, vol. 11, no. 1, pp. 72-81, 2015.
[4] C.C. Chinedu and Y. Kamin, "Strategies for improving higher order thinking skill in teaching and learning of design and technology education," *Journal of Technical Education and Training*, vol. 7, no. 2, pp. 35-43, 2015.
[5] S Wang and H Wang, "Teaching and learning higher-order thinking," *International Journal of Arts & Sciences,* vol. 7, no. 2, pp. 179-189, 2014.
[6] A. Z. Abidin, E. Istiyono, N. Fadilah, and W. S. B. Dwandaru, "A Computerized adaptive test for measuring the physics critical thinking skills in high school students," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 8, no. 3, pp. 376-383, 2019.
[7] M.D. Kusuma, U. Rosidin, Abdurahman, and A. Suyatna, "The Development of Higher Order Thinking Skill (HOTS) instrument Assessment in Physics Study," *Journal of Research and Method in Education*, vol. 7, no. 1, pp. 26-32, 2017.
[8] B. P. Mainali, "Higher order thinking in education," *A Multidisciplinary Journal,* vol. 2 no, 1, pp. 5-10, 2012.
[9] K. Wardany, Sajidan, and R. Murni, "Pengembangan penilaian untuk mengukur higher order thinking siswa," *Jurnal Inkuir,* vol. 6, no. 2, pp. 1-16, 2017.
[10] L. W. Anderson and D. R. Krtahwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy*. New York: Longman Publishing, 2001.
[11] G. V. Madhuri, V. S. S. N. Kantamreddi, and L. N. S. Prakash Goteti, "Promoting higher order thinking skills using inquiry-based learning," *European Journal of Engineering Education*, vol. 37, no. 2, pp. 117-123, 2012.
[12] J. W. Mahoney and B. Harris-Reeves, "The effects of collaborative testing on higher order thinking: Do the brightget brighter?" *Active Learning in Higher Education*, vol. 20, no. 1, pp. 25-37, 2019.
[13] Y. M. Heong, *et al.,* "The Level Marzano Higher Order Thinking Skills Among Technical Education Students," *International Journal of Social Science and Humanity,* vol. 1, no. 2, pp. 121-125, 2011.
[14] B. Limbach and W. Waugh, "Developing higher level thinking," *Journal of Instructional Pedagogies,* vol. 3, no. 1, pp. 1-9, 2010.
[15] C. Fischer, L. Bol, and S. Pribesh, "An investigation of higher-order thinking skills in smaller learning community social studies classrooms," *American Secondary Education,* vol. 39, no. 2, pp. 5-26, 2011.
[16] A. B. Nordin and N. Alias, "Learning outcomes and student perceptions in using of blended learning in history," *Procedia - Social and Behavioral Sciences*, vol. 103, pp. 577–585, 2013.
[17] R.A. Kurniawan and D. Lestari, "The Development Assessment Instruments of Higher Order Thinking Skills on Ekonomc Subject," *Dinamika Pendidikan*, vol. 14, no. 1, 102-115, 2019.
[18] A. Y. Gunduz, E. Alemdag, S. Yasar, and M. Erdem, "Design of a problem-based online learning environment and evaluation of its effectiveness," *The Turkish Online Journal of Educational Technology*, vol. 15, no. 3, pp. 49-57, 2016.
[19] A. C. Saputri, S. Sajidan, Y. Rinanto, A. Afandi, and N. M. Prasetyanti, "Improving students' critical thinking skills in cell-metabolism learning using stimulating higher order thinking skills model," *International Journal of Instruction*, vol. 12, no. 1, pp. 327-342, 2018.
[20] P. Afflerbach, B. Y. Cho, and J. Y. Kim, "Conceptualizing and assessing higher-order thinking in reading," *Theory into Practice*, vol. 54, no. 3, pp. 203-212, 2015.
[21] W. J. Boone, R. J. Staver, and S. M. Yale, *Rasch Analysis in the Human Sciences*. London: Springer, 2014.
[22] S. M. Haley, W. J. Coster, H. M. Dumas, M. A. Fragala-Pinkham, J. Kramer, P. Ni, *et al.*, "Accuracy and precision of the Pediatric Evaluation of Disability Inventory computer-adaptive tests (PEDI-CAT)," *Developmental Medicine and Child Neurology*, vol. 53, no. 12, pp. 1100-1106, 2011.
[23] M. Yuniar, C. Rakhmat, and A. Saepulrohman, "The Analyses of HOTS (High Order Thinking Skills) in Objective Test in Social Studies Class 5 th SD Negeri 7 Ciamis," (in Bahasa), *Jurnal Ilmiah Mahasiswa Pendidikan Guru Sekolah Dasar*, vol. 2, no. 2, pp.187-195, 2015.
[24] Pi'i, "Developing Higher-Level Thinking Learning and Assessment in History Subjects," (in Bahasa), *Sejarah dan Budaya*, vol. 10, no. 2, pp.197-208, 2016, doi: 10.17977/um020v10i22016p197.
[25] S. Avargil, O. Herscovitz, and Y. J. Dori, "Teaching thinking skills in context-based learning: Teachers' challenges and assessment knowledge," *J Sci Educ Technol*, vol. 21, pp. 207–225, 2012, doi: 10.1007/s10956-011-9302-7
[26] D. Drake, D. Frederick, and L. R. Nelson, *Engagement in teaching history theory and secondary teachers*. New Jersey: Pearson, 2005.
[27] N. Sener, C. Turk, and E. Tas, "Improving science attitude and creative thinking through science education project: A design, implementation and assessment," *Journal of Education and Training Studies*, vol. 3, no. 4, pp. 57-67, 2015.
[28] G.T. Dam and M. Volman, "Critical thinking as a citizenship competence: teaching strategies," *Learning and instruction,* vol. 14, no. 4, pp. 359-379, 2004.
[29] D. Andrian, B. Kartowagiran, and S. Hadi, "The instrument development to evaluate local curriculum in Indonesia," *International Journal of Instruction*, vol. 11, no. 4, pp. 922-934, 2018.
[30] F.A. Karim and M. Puteh, "The development of higher order thinking skill assessment instrument for word problem," *International Journal of Academic Research in Business and Social Sciences*, vol. 9, no. 6, pp. 1097-1083, 2019.
[31] S. Ramadhan, D. Mardapi, Z. K. Prasetyo, and H. B. Utomo, "The development of an instrument to measure the higher order thinking skill in physics," *European Journal of Educational Research,* vol. 8, no. 3, pp. 743-751, 2019.
[32] W. R. Borg, and J. P. Gall, *Education research*. New York: Allyn and Bacon, 2003.
[33] I. Kinay and B. Bagceci, "The investigation of the effects of authentic assessment approach on prospective teachers' problem-solving skills," *International Education Studies*, vol. 9, no. 8, pp. 51-59, 2016.

[34] D. J. Weiss, "Better data from better measurements using computerized adaptive testing," *Journal of Methods and Measurement in the Social Sciences,* vol. 2, no. 1, pp. 1-27, 2011.
[35] Aman, "Final Examination Test Instruments for History Subject in Yogyakarta, Indonesia: A Quality Analysis," *Universal Journal of Educational Research,* vol. 7, no. 12, pp. 2857-2866, 2019.
[36] Z. Arifin, *Learning Evaluation*, (in Bahasa). Bandung: PT Remaja Rosdakarya, 2016.
[37] C. Boopathiraj and K. Chellamani, "Analysis of test items on difficulty level and discrimination index in the test for research in education," *International Journal of Social Science & Interdisciplinary Research*, vol. 2, no. 2, pp.183-193, 2013.
[38] G. T. L. Brown, S. E. Irving, and P. J. Keegan, *An Introduction to educational Assessment, Measurement & Evaluation*. Auckland: Dunmore Publishing, 2014.
[39] M. Pommerich, "Developing computerized versions of paper-and pencil tests: Mode effects for passage-based tests," *The Journal of Technology, Learning and Assessment,* vol. 2, no. 6, pp. 3-44, 2004.
[40] J. Szilagyi, D. H. Clements, and J. Sarama, "Young Children's Understandings of Length Measurement: Evaluating a Learning Trajectory," *Journal for Research in Mathematics Education,* vol. 44, no. 3, pp. 581-620, 2013.
[41] K. Coaley, *An Introduction to Psychological Assessment and Psychometrics*. London: SAGE Publications Inc, 2010.
[42] A. M. L. Cavallo, W. H. Potter, and M. Rozman, "Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors," *School Science and Mathematics*, vol. 104, no. 6, pp. 288-300, 2004.
[43] D. F. Polit and C. T. Beck, "The content validity index: are you sure you know what's being reported? Critique and recommendations," *Research in Nursing & Health,* vol. 29, no. 5, pp. 489-49, 2006.
[44] J. Setiawan, Aman, and Wulandari, "Understanding Indonesian history, interest in learning history and national insight with nationalism attitude," *International Journal of Evaluation and Research in Education (IJERE),* vol. 9, no. 1, pp. 1-10, 2020.
[45] C. A. Wynd, B. Schmidt, and M. A. Schaefer, "Two quantitative approaches for estimating content validity," *Western Journal of Nursing Research*, vol. 25, no. 5, pp. 508–518, 2003.
[46] S. Azwar, *Validity and Reliability*, (in Bahasa). Yogyakarta, Indonesia: Pustaka Pelajar, 2007.
[47] D. Mardapi, *Techniques for drafting test and non-test instruments*, (in Bahasa). Yogyakarta, Indonesia: Mitra Cendekia, 2008.
[48] H. Retnawati, *Validity, reliability and item characteristics*, (in Bahasa). Yogyakarta, Indonesia: Parama Publishing, 2016.
[49] A. Sudijono, *Introduction to Evaluation*, (in Bahasa). Jakarta: Rajawali Pers, 2012.
[50] D. Mardapi, *Measurement of educational assessments and evaluations*, (in Bahasa). Yogyakarta, Indonesia: Nuha Medika, 2012.
[51] M. Arif, "Application of Anates Application Form multiple choice questions," (in Bahasa), *Jurnal Ilmiah Edutic*, vol. 1, no.1, pp. 1-9, 2014.
[52] J. E. Barnett and A. L. Francis, "Using higher order thinking questions to foster critical thinking: A classroom study," *Educational Psychology,* vol. 32, no. 2, pp 201-211, 2012.