JOURNAL OF BIOLOGY EDUCATION
J O B E

E-ISSN 2656-3436/ P-ISSN 2615-3947
IAIN KUDUS
http://journal.iainkudus.ac.id/index.php/jbe

The Development of High Order Thinking Skills Test based on Google Forms on Cell Biology Materials

***Setiyo Prajoko[1], Alfiyah Erawati[2], Fiki Wafara Amali[3], Meifiqih Zunaena[4], Indria Arganingtias[5]**

[1) 2) 3) 4) 5)] Departement of Biology Education, Faculty of Teacher Training and Education, Universitas Tidar
[1] setiyoprajoko@untidar.ac.id, [2] alfierawati27@gmail.com, [3] fikiwafaraamali@gmail.com, [4] meifiqih.zunaena21@gmail.com, [5] indriaarganingtias02@gmail.com

***correspondence author**

ABSTRACT

This study aims to develop a google form-based evaluation instrument to improve students' High Order Thinking Skills in Cell Biology materials. The research and development (R & R&D) method that uses a 4-D development design (Define, Design, Develop, Disseminate). This research was conducted from May to June 2021 in senior high schools in the Magelang district. The subjects in this study were 50 students of class XI-MIPA. Based on the observation data, the validity level is declared valid in terms of the validity level. The data obtained from this reliability meet the reliability criteria in the good category. The difficulty level is moderate and has good quality questions from the obtained data. From the data received, the discriminatory power of the questions that have been carried out can be found in the categories of excellent, good, sufficient and bad questions. From the data obtained based on the construct validity test, it can be seen that the test is included in the excellent category.

Keywords: HOTS, cell biology, evaluation, multiple-choice test

INTRODUCTION

The development of science and technology in the 21st century is very rapid. Implementing education, especially science education, must also adapt to these developments. To be competitive in this century, several skills need to be mastered by students. Based on The Partnership for 21st-century learning (2015), there are three frameworks of skills that students need to master, namely life and career skills (life and career skills), innovation and learning skills (learning and innovation skills), and media information, and technology skills. (Information, media, and technology skills). Griffin et al. (2012) reveal a need to change the education system in the 21st century. Change is aimed at learning through digital networks and collaborative problem-solving. Ananiadoui & Claro (2009) revealed that critical success factors for 21^{st}-century learning policies include quality and relevant teacher training, curriculum integration, and transparent and appropriate assessments (S Prajoko, Amin, Rohman, & Gipayana, 2016).

Advances in information technology now have a lot of positive impacts on progress in the world of education, especially in computer technology and internet technology, both in the form of hardware and software, providing many offers and choices for the world of education in supporting the learning process (Rolisca & Achadiyah, 2014). Facing these events, the world of education must always be ready to adapt technological developments to improve the quality of education, especially adjustments in the learning process in schools. In the learning process, there are three components, namely objectives, learning activities, and evaluation (Laelasari & Hilmi Adisendjaja, 2018). Assessment must provide comprehensive information that helps teachers improve their teaching abilities and help students achieve optimal educational development (Wahyuningsih, Wahyuni, & Lesmono, 2016).

Learning evaluation is an assessment process to make a decision based on the results of a comprehensive assessment, including; affective aspects (attitudes), cognitive aspects (knowledge), and psychomotor aspects (skills). For this reason, the evaluation instrument used should be able to provide comprehensive measurement and assessment results, covering all aspects of student learning outcomes (Sanjaya, Asyhar, & Hariyadi, 2015).

The importance of mastering higher-order thinking skills is contained in several points of the Competency Standards for High School Graduates. Lewis & Smith (1993) define higher-order thinking skills (The Higher Order Thinking Skills) as thinking skills that occur when a person takes new information and information already stored in his memory, then relates the information and conveys it to achieve the goals or answers needed.

HOTS can be interpreted as the ability of complex thinking processes that include parsing material, criticizing and creating solutions to problem-solving (Laelasari & Anggraeni, 2017., (Budiarta, 2021). Responding to the same thing, Thomas & Thorne (2009) define HOTS as the ability to think by making connections between facts and problems. Problem-solving is done not only through the process of remembering or memorizing but requires making connections and conclusions from issues. Accompanying similar things, HOTS is the ability to combine facts and ideas in the process of analyzing, evaluating to the stage of creating in the form of providing an assessment of a fact that is learned or being able to start from something that has been studied (Annuuru, Johan, & Ali, 2017).

Analyzing, evaluating and creating is part of the cognitive taxonomy created by Benjamin S. Bloom in 1956. In the end, it was refined by Anderson and Krathwohl (2001) into C1-remembering, C2-understanding, C3- applying, C4-analyzing, C5-evaluating, and C6-creating. Tanujaya et al. (2017) explain that levels one to three are low-level thinking skills or LOTS (Lower Order Thinking Skills), and levels four to six are HOTS (Higher Order Thinking Skills). So, when viewed from the cognitive domain, HOTS is the ability to analyze, evaluate and create. Based on Sulianto et al., 2018) pre-present picture of the mental level in the revised Bloom's taxonomy in Figure 1.

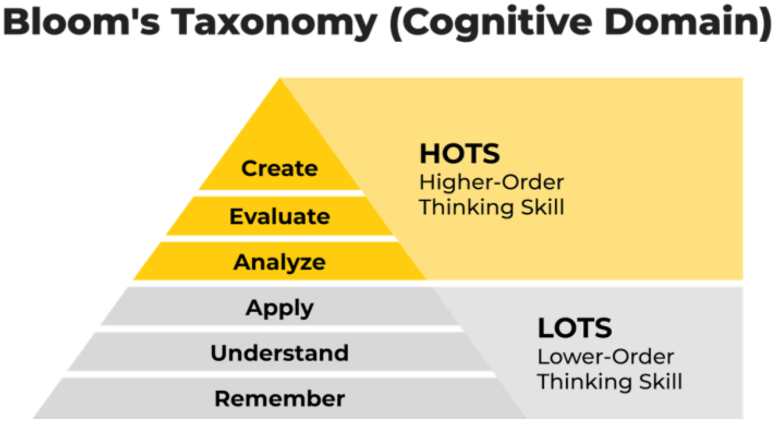

Figure 1 The Position of High Order Thinking Skills in Bloom's Taxonomy

Based on Figure 1, the processes of analyzing and evaluating as critical thinking; At the same time, creation is part of creative thinking skills; critical and creative thinking process abilities are used to solve problems or create solutions to make decisions. The three cognitive processes are moved when they find new issues. The success of higher-order thinking skills is located in a person's success in moving the three thinking processes (Saido, Siraj, Nordin, & Al_Amedy, 2018).

According to Arifin (2009), a teacher can develop a form of test that contains both objective and subjective questions. An example of a factual question is a multiple-choice test. Multiple-choice tests are usually used to make tests that cover broad learning objectives, are quick and easy to correct, can eliminate subjectivity in correction, and can diagnose difficulties in student learning outcomes. At this time, many teachers still use formative tests using written tests that allow students to dishonestly cheat during tests, especially for multiple-choice test questions (Wahyuningsih et al., 2016).

When working on multiple-choice test questions, students' cheating is expected to be overcome using an online exam system. The online exam system of multiple-choice test questions allows students to take the test honestly to create high order thinking skills. In the online exam system for multiple-choice test questions, the teacher can determine the time limit for working on the questions and design a package of questions randomly so that one student has different questions with the same number of questions. In addition, the online test system for multiple-choice test questions has other advantages, including having statistical features, analysis of results, flexible data results, complete display settings so that it can support various types of tests and is easy to use and manufacture mainly by using the google form (Utomo, 2015).

The familiarity of society at this time with various technological products such as computers, tablets and smartphones and the availability of increasingly cheap internet connections are also opportunities for the use of information and communication technology in the implementation of the education system by using one of the software that is easily accessible, free to use, simple in operation, and good enough to be developed as a tool for evaluating the performance of lecturers in the learning process, namely Google Form. Google Forms is a component of the Google Docs service. This application is perfect for students, teachers, lecturers, office employees, and professionals who like to make quizzes, forms, and online surveys. Features of Google Forms can be shared openly or specifically to Google account owners with various accessibility options, such as read-only (can only read) or editable (can edit documents). In addition, Google doc is an alternative for people who don't have the funds to buy paid applications, so most people prefer to use free programs rather than hijack paid programs like Microsoft Office because pirating the program is not good (Batubara, 2016; Sahlani & Agung, 2020).

As for some of the functions of the Google Form for online learning, namely providing online practice questions/tests to students through the website page, collecting various opinions from other people through the website page, and managing various student/teacher data through the website page. Based on the above background, this study aims to distribute questionnaires in biology questions to students via google form and measure the high-level thinking of 11th-grade high school students in the Magelang district on cell biology material (Batubara, 2016).

Making evaluation tools in the learning process using Google Forms allows teachers to assess students in the learning process from anywhere and anytime, as long as they have a computer, laptop, or cellphone connected to the internet. In addition, educators will feel helped by Google Forms' ability to recapitulate the assessment results and present them in presentations that can be analyzed and presented as desired (Batubara, 2016).

According to Fauzi (2014) research results reveal that the use of Google Forms as a learning evaluation tool in biology subjects starts from the planning stage, the readiness of facilities and infrastructure, development Google Forms, to the implementation stage of using Google Forms in learning evaluation activities has an impact and benefit both from the practical aspect. Efficiency, attractiveness, and appearance design. For educators, the Google

Form is greatly helped in terms of cost, time, and effort. For students themselves to be more interested

RESEARCH METHODS

This research and development (R&D) used the 4-D development design by Thiagarajan (1974). Research and Development (R&D) is an effort or activity to develop an effective product for use by schools and not to test the theory. The 4-D development model is a learning device development model. This research was carried out in May-June 2021 at a high school in the Magelang district. The subjects in this study were 50 students of an eleven-grade science class.

1. Defining Stage

The Defining stage, which includes front-end analysis, student analysis, task analysis, concept analysis, and learning objectives, is then used to formulate test questions. Front-end analysis aims to emerge and define the fundamental problems encountered in learning (Thiagarajan, 1974). In this development research, the material specified is "cell biology".

2. Designing stage

The Design stage aims to produce a high-level thinking ability test design for cell biology learning in high school students. The design phase includes preparing learning outcomes tests based on students' high-level thinking skills on cell biology, media selection, format selection, and initial design. The subject of cell biology consists of the history of the discovery of cells, the structure and function of cell parts, and the differences between animal and plant cells. After that, a grid of questions and scoring guidelines was made. At the end of this stage, a design in a multiple-choice test was obtained, totaling 50 questions (Nugraha & Widiyaningrum, 2015).

3. Developing stage

In the Development stage, validity tests were carried out, including content validity and construct validity. Content validity test includes validity test, reliability test, discriminatory power test, and difficulty index. The content validity test asked 50 high school students in Magelang Regency who had finished taking cell biology subjects. The data analysis technique for construct validity was calculated using the equation:

Validator average of al aspets: $Va = \frac{\sum_{i=1}^{n} A_i}{n}$

Data analysis techniques for validity with the concurrent validity method:

$$r_{xy} = \frac{\Sigma_{xy}}{(\Sigma_{x^2})(\Sigma_{y^2})}$$

The technique of analyzing item validity data uses the formula:

$$\gamma pbi = \frac{M_p - M_t}{S_t}\sqrt{\frac{p}{q}}$$

note:
γpbi : biserial correlation coefficient
Mp: the average score of the subjects who answered correctly for the item whose validity was sought
Mt: mean total score
St: standard deviation of the total score
p: the proportion of students who answered correctly
q: the proportion of students who answered incorrectly.

The reliability data analysis technique uses the Spearman-Brown formula. The formula used to calculate reliability is the KR 11 Spearman-Brown formula as follows:

$$\mathrm{r}_{11} : \frac{\mathrm{n}}{(n-1)}\,(\frac{\mathrm{M}(\mathrm{n}-\mathrm{m})}{nSt\ 2})$$

note:
r_{11} = instrument reliability
n = number of questions
M = score average
S2/t = total variance

Data analysis techniques for the level of difficulty of the questions:

$$P = \frac{B}{JS}$$

note:
P = difficulty index
B = the number of students who answered the question correctly
JS = total number of students taking the test

My data analysis technique differentiates the questions:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B$$

note:
J = number of test-takers
JA = number of participants in the upper group
JB = number of lower group participants
BA = the number of participants in the upper group who answered the question correctly
JB = the number of participants in the lower group who responded to the question correctly.

The validity criteria for determining the level of validity of the high order thinking test instrument in e-learning-based biology learning are according to Table 1.

Table 1. The criteria of validity

Criteria	Interval
invalid	$1 \leq V < 2$
less valid	$2 \leq V < 3$
quite valid	$3 \leq V < 4$
valid	$4 \leq V < 5$
very valid	$= 5$

(Source: Sari, 2018)

The results of the calculation of discriminatory power are classified as poor, adequate, good, and very good. Sudijono (2001) ranks the discriminating power of question items as presented in table 2.

Table 2. The interpretation of the discriminating power

Interval	Criteria	Interpretation
$P < 0{,}20$	bad	the discriminating power is bad and needs revision
$0{,}20 \leq a < 0{,}40$	enough	the discriminating power is enough no need for revision
$0{,}40 \leq a < 0{,}70$	good	the discriminating power is good no need for revision
$0{,}70 \leq a < 1{,}00$	excellent	the discriminating power is excellent no need for revision

(Sudijono, 2001)

Calculating the difficulty index of the item items is classified as easy, medium, and difficult (Arikunto, 2021). Ranks the difficulty index presented in Table 3.

Table 3. The criteria of difficulty index

Difficulty index	criteria
t > 0.70	easy
0.30 ≤ t ≤ 0.70	medium
t<0.30	difficult

The technique of analyzing higher-order thinking skills is the maximum score is the highest score is 100. The criteria for students' higher-order thinking abilities are in Table 4.

Table 4. The criteria for higher-order thinking skills

Score	**Criteria**
80-100	excellent
66-79	good
56-65	enough
40-54	lack
0-39	very lack

(Source: Lewy et al., 2009)

The results of the content validity test were then evaluated and revised. After that, a construction validity test was carried out by one of the lecturers of the biology education study program to test the item questions with aspects—measurement of high order thinking skills based on google form on cell biology material. Construct validity was tested for instrument format, content, question construction, and language.

The assessment of this validator follows a Likert scale of 1-4 for each assessment item (Arsyad, 2007). The average assessment results are classified according to the criteria presented in Table 5.

Table 5. The criteria of construct validity

Score average	Criteria
3,5 ≤ M ≤ 4,0	Very good
2,5 ≤ M ≤ 3,5	Good
1,5 ≤ M ≤ 2,5	Enough
M < 1,5	Bad

(source: Nurdin, 2007)

4. Disseminating stage

Dissemination activities are aimed at distributing the information of research results to society. Research results in the form of assessment products are distributed in a limited environment with biology teachers in schools in which research was conducted. Besides that, dissemination is also carried out by providing information to biology teachers at the Magelang Regency Senior High School. They are members of the MGMP (an association of biology teachers in Magelang).

RESULT AND DISCUSSION

The results of the development research obtained are tests measuring high order thinking skills based on google form on cell biology material. Based on the flow of the 4D development model with the research objective, namely to determine the validity based on expert judgment, to assess the quality of the HOTS test trial in terms of validity, reliability, level of difficulty, and distinguishing power on cell biology material for eleven-grade students. Based on the formulation of the problem and the flow of the research, the following results are obtained.

Initial Draft Products

The initial draft product was developed by conducting a test grid based on core competencies, essential competencies, and learning indicators. Then, set it into the form of HOTS-based multiple-choice questions. The research was conducted on XI high school students. Bad tests should be discarded or not used to score students. A test can be good as a measuring tool if it meets the test requirements. The requirements for a good test are valid, reliable, and have distinguishing power and a good level of difficulty. The essential test requirement is validity. A test is called reasonable if the test can accurately measure what is being measured. The good or bad of a test or evaluation tool can be viewed from its validity, reliability, level of difficulty and distinguishing power (Nuswowati, Binadja, & Ifada, 2010).

The results of this study include content validity and construct validity tests. Content validity has item validity, reliability, discriminating power, and difficulty index. Construct validity consists of the material domain, the construction domain, and the language domain.

Content Validity

The trial in this study was carried out on eleven-grade high school students with a total sample of 50 students. The analysis of the validity of the items was carried out with Microsoft Excel 2016, and the results of the calculation of the truth of the questions can be seen in Table 6 below.

Table 6. The Recapitulation of the validity test result

aspects	valid	invalid
question number	2, 6, 7, 8, 11, 12, 13, 14, 15, 16, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 44, 45, 46, 47, 49, 50	1, 3, 4, 5, 9, 10, 17, 18, 26, 41, 43, 48
TOTAL	38 (76%)	12 (24%)

A test is called valid or valid if it can accurately measure what it is intended to measure. The percentage of valid questions is 76%, and the rate of invalid questions is 24%. The test items on questions 1, 3, 4, 5, 9, 10, 17, 18, 26, 41, 43, and 48 are bad, so they have low or invalid validity. This can be seen if calculating the discriminatory power and difficulty level also gets bad results (Setiyo Prajoko, Anjani, Oktaviani, Fathimah, & Kamaludin, 2021; Solichin, 2017).

Items that are called invalid should be corrected, and things that are called valid can be reused. Questions that have been declared valid must be maintained by documenting the questions in the question bank. Invalid items should be corrected by increasing the mastery of the researcher's technique in preparing the articles. Based on the item validity data, it can be concluded that the questions in this study are pretty good in terms of their level of validity (Marthunis et al., 2015)

Reliability

The trial in this study was carried out on eleven-grade high school students with a total sample of 50 students. The analysis of the validity of the items was carried out with Microsoft Excel 2016. The results of the calculation of the validity of the questions can be seen in the table below.

Table 7. The Result of the Reliability test

Cronbach alpha (α)	**Category**
0,94	Reliable

Based on Table 7, it can be seen that the reliability testing of the instrument being tested is reliable, which means that the instrument whose measurement results are made can be trusted because when tested repeatedly, it gives fixed measurement results, where this is obtained from external reliability testing with the Spearman-Brown formula. This is based on the calculation of the alpha (α) score obtained by 0.94, while the R-Table score with a deviation of 5% is 0.44, so the score is greater than the R-table score. A test score is reliable if the alpha score is greater than the R-Table score (Lestari et al., 2016).

According to Suryabrata (2000), reliability shows how much the measurement results with this tool can be trusted. The measurement results must be reliable, which means they must have consistency and stability. This shows that the evaluation instrument made has a category of reliability.

Difficulty Level

After going through the level of difficulty testing, the results are shown in Table 8.

Table 8. The Recapitulation of the difficulty index result

Category	Question Number	TOTAL
accessible	1, 2, 3, 5, 7, 8, 9, 15, 38, 43, 49	11 (22%)
medium	4, 6, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 44, 45, 46, 47, 48, 50	38 (76%)
difficult	14	1 (2%)

The items in the table have different levels of difficulty adjusted to the group to be achieved according to the learning objectives. The difficulty level analysis data showed that the types of questions were the accessible (22%), medium (76%), and complex (2%). Item number 1 has an easy question level, but it is tested on an invalid validity analysis; this can be caused because the questions are considered too easy. The goals achieved are still not suitable for questions that are too easy. Question number 26 has a high level of difficulty, but the question has validity, so the question still has a category that can be used. Good questions are items that are not too difficult and not too easy, so a good question is a question that has a moderate level of difficulty (Arikunto, 2021)

Difference Power

The discriminatory power of a question is distinguishing between intelligent students (high capacity) and less intelligent students (low ability). This can be seen in the table of which items are lacking in determining the purpose of distinguishing power (Solichin, 2017: 197). The result of the distinguishing power test.

Table 9. The result of the distinguishing power test

Category	Question Number	TOTAL
bad	1, 5, 48	3 (6%)
medium	2, 3, 4, 6, 7, 8, 9, 10, 11, 14, 15, 17, 18, 19, 23, 26, 30, 37, 38, 43, 45, 50	22 (44%)
good	13, 16, 20, 21, 22, 24, 25, 27, 28, 29, 31, 32, 33, 34, 35, 36, 39, 40, 41, 42, 44, 46, 47, 49	24 (48%)
excellent	12	1 (2%)

In the table, it can be seen that item numbers 1, 5, and 48 have poor distinguishing power after testing. The discrimination index generated on the question only slightly results in insufficient discriminating power. Later, it cannot achieve the goal of distinguishing the level of student ability.

The purpose of the discriminatory analysis is to examine the ability of students' questions between high achievement students and low achievement students. A good question will be solved well in a group of intelligent students, but if the question is given to a group of students who are less, the result is bad. On questions that do not have good discriminatory power, if the questions are given to intelligent students and children lacking, they will provide the same results. Based on biased analysis data, it shows that the characteristics of the questions are in a wrong category (6%), adequate (44%), good (48%), and excellent (2%). From these results, the questions developed are homogeneous.

Based on the data analysis of the discriminatory power of the questions that have been carried out, it is found that the categories of questions are good, sufficient and bad. Good discrimination power is not too easy and not too difficult or moderate; items with a bad discrimination index can be immediately discarded or not used (Alwi, 2015; Pangestuti, Febriyana, Adhawiyah, Febriyanti, & Prajoko, 2021). According to Arikunto (2021), several reasons that the items have low or poor discriminating power are, among others, questions that contain bias, questions that are too difficult, and unreasonable distractors.

Construct Validity

Analysis of the validity of the questions was carried out based on the assessment by the validator. The results of the assessment can be seen in table 10.

Table 10. Recap of Question Validation Results

Aspects	$\bar{x}$	Criteria
Materials	3,44	good
Construction	4	very good
Language	4	very good
TOTAL AVERAGE	3,81	very good

The table shows that the developed test is included in the excellent category with a score of 3.81, so that the instrument meets the criteria very well and can be tested.

Questions are declared valid or have high validity, namely questions that can measure the expected competence. At the same time, the questions that are invalid or have low validity are those that cannot measure the desired competence. A good question is a question that can be tested (Rahmani, Ningsih, & Nurdini, 2015).

The instruments that have been tested have generally met the criteria for content validity. Of the fifty HOTS items tested, 38 of them met content validity. Only item items in numbers 1, 3, and 5 did not meet the overall content validity, the Pearson correlation validity test, the easy difficulty index, and poor discriminatory power. After reviewing the characteristics of questions number 1, 3, and 5, they are burdened with concepts and do not refer to the criteria for the HOTS questions.

In fact, according to Budiarta's (2021) explanation, HOTS can be interpreted as the ability of complex thinking processes that include parsing material, criticizing and creating solutions to problem-solving. Therefore, the items numbered 1, 3, and 5 need a total revision before being used for research activities. Meanwhile, item number 14 meets the criteria for the validity of the Pearson correlation and the difficulty index, which is challenging but has poor or sufficient discriminatory power. Thus, item number 14 needs to be revised slightly. Revision of item questions is done by reviewing the sentence structure and language and whether there is an element of ambiguity. Dillashaw & Okey (1980) revealed that item items that we cannot distinguish between students who got high scores and students who got low scores needed to be revised by simplifying sentences and making item items easier to read and understand by students. Questions that are too easy also need to be replaced by improving the quality of questions that reflect higher-order thinking skills.

The content validity results obtained are relevant to the effects of research conducted by Sari et al. (2018), students and lecturers of Yogyakarta State University. The results showed that the test questions for a junior school in Gunungkidul Regency had excellent content validity, with 38 items meeting all aspects of the study sheet and 12 items not meeting 100% of the overall review criteria. The twelve items that were declared unfavorable were items that did not meet the material and construction aspects. In the material element, the questions according to the indicators and construction aspects, the questions whose answer choices are homogeneous and the length of the answer choices are approximately the same (Sanjaya et al., 2015).

Based on Table 10, it can be seen that the acquisition of construct validation carried out by experts as validators are included in the outstanding category. Aspects of the construct in this study include material, construction, and language. The item questions are also by the indicators developed by integrating HOTS questions on cell biology material and the purpose of learning biology on cell biology material in eleven-grade.

The instrument uses clear item numbering, the type and size of the letters follow the standards, the layout and layout of the writing are by the guidelines for writing multiple-choice questions, the contents of the instrument are by the measured aspects, and the material on the questions has been formulated.

The language aspect is by good and correct Indonesian rules. Writing questions have used terms that are easy to understand. Based on Suwarto (2017), revealed that the use of sound and correct Indonesian affects determining construct validity. The expert gave several notes to use more straightforward language in writing the scoring guidelines so that they are easy to understand.

CONCLUSION

Based on the research results above, it can be concluded that the content validity test is based on the validity of the items included in the valid (76%), invalid (24%) category and included in the reliable category with a score of 0.94. The results of the content test validity indicate that the test kit is feasible to use and apply. The content validity test also has a quality level of difficulty: easy (22%), moderate (76%), and complex (2%) categories with discriminatory power: poor type (6%), adequate (44%), good (48%), and perfect (2%). Distinguishing power on the test device has an excellent high score; the questions vary to measure students' abilities. The construct validity results indicate that the developed test is included in the excellent category with a score of 3.81 so that the instrument meets the criteria and can be tested.

REFERENCES

Alwi, I. (2015). Kriteria empirik dalam menentukan ukuran sampel pada pengujian hipotesis statistika dan analisis butir. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, *2*(2).

Ananiadoui, K., & Claro, M. (2009). *21st century skills and competences for new millennium learners in OECD countries*.

Annuuru, T. A., Johan, R. C., & Ali, M. (2017). Peningkatan kemampuan berpikir tingkat tinggi dalam pelajaran ilmu pengetahuan alam peserta didik sekolah dasar melalui model pembelajaran treffinger. *Educational Technologia*, *1*(2).

Arifin, Z. (2009). *Evaluasi pembelajaran* (Vol. 118). Bandung: PT Remaja Rosdakarya.

Arikunto, S. (2021). *Dasar-Dasar Evaluasi Pendidikan Edisi 3*. Bumi Aksara.

Arsyad, N. (2007). Model Pembelajaran Matematika yang Menumbuhkan Kemampuan Metakognisi untuk Menguasai Bahan Ajar. *Disertasi. Tidak Diterbitkan. Surabaya: Universitas Negeri Surabaya*.

Batubara, H. H. (2016). Penggunaan google form sebagai alat penilaian kinerja dosen di Prodi PGMI Uniska Muhammad Arsyad Al Banjari. *Al-Bidayah: Jurnal Pendidikan Dasar Islam*, *8*(1).

Budiarta, L. G. R. (2021). Developing HOTS-Based Students' Worksheet for Fifth-Grade Elementary School. *Journal of English Education and Teaching*, *5*(4), 468–488.

Dillashaw, F. G., & Okey, J. R. (1980). *A Test of the Integrated Science Process Skills for Secondary Science Students.*

Fauzi, M. R. (2014). *Penggunaan Google Form Sebagai Alat Evaluasi Pembelajaran Pada Mata Pelajaran Bahasa Indonesia: Studi Deskriptif Analitis pada Kelas VIII di Sekolah Menengah Pertama Negeri 1 Lembang*. Universitas Pendidikan Indonesia.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In *Assessment and teaching of 21st century skills* (pp. 1–15). Springer.

Laelasari, I., & Anggraeni, S. (2017). Improving Critical Thinking and Metacognition Ability Using Vee Diagram through Problem-Based Learning of Human Respiratory System. *Atlantis Press*, 45–51. https://doi.org/10.2991/icmsed-16.2017.16

Laelasari, I., & Hilmi Adisendjaja, Y. (2018). *Mengeksplorasi Kemampuan Berpikir Kritis Dan Rasa Ingin Tahu Siswa Melalui Kegiatan Laboratorium Inquiry Sederhana* (Vol. 01). Retrieved from http://journal.stainkudus.ac.id/index.php/Thabiea

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory into Practice*, *32*(3), 131–137.

Lewy, L., Zulkardi, Z., & Aisyah, N. (2009). Pengembangan soal untuk mengukur kemampuan berpikir tingkat tinggi pokok bahasan barisan dan deret bilangan di kelas IX akselerasi SMP Xaverius Maria Palembang. *Jurnal Pendidikan Matematika*, *3*(2), 14–28.

Nugraha, M. I., & Widiyaningrum, P. (2015). Efektivitas Media Scratch Pada Pembelajaran Biologi Materi Sel di SMA Teuku Umar Semarang. *Journal of Biology Education*, *4*(2).

Nuswowati, M., Binadja, A., & Ifada, K. E. N. (2010). Pengaruh validitas dan reliabilitas butir soal ulangan akhir semester bidang studi kimia terhadap pencapaian kompetensi. *Jurnal Inovasi Pendidikan Kimia*, *4*(1).

Pangestuti, D., Febriyana, D., Adhawiyah, I. R., Febriyanti, R. T., & Prajoko, S. (2021). Development of Quizizz Application-Based Test to Measure Science Process Skills of High School Students on Biodiversity Materials. *Indonesian Journal of Biology Education*, *4*(1), 36–44.

Prajoko, S, Amin, M., Rohman, F., & Gipayana, M. (2016). Pengembangan Tes Keterampilan Proses Sains Pada Mata Kuliah Praktikum IPA di SD. *Seminar Nasional IPA*, *7*(2), 1.

Prajoko, Setiyo, Anjani, A., Oktaviani, R. D., Fathimah, P., & Kamaludin, W. (2021). Development of Self-assessment Instruments to Measure Student Affective Domains on Online Biology Learning. *THABIEA: JOURNAL OF NATURAL SCIENCE TEACHING*, *4*(2), 185–197.

Rahmani, M., Ningsih, K., & Nurdini, A. (2015). Analisis Kualitas Butir Soal Buatan Guru Biologi Kelas X SMA Negeri 1 Tanah Pinoh. *Jurnal Pendidikan Dan Pembelajaran*

Khatulistiwa, *4*(2).

Rolisca, R. U. C., & Achadiyah, B. N. (2014). Pengembangan media evaluasi pembelajaran dalam bentuk online berbasis e-learning menggunakan software wondershare quiz creator dalam mata pelajaran akuntansi SMA Brawijaya Smart School (BSS). *Jurnal Pendidikan Akuntansi Indonesia*, *12*(2).

Sahlani, L., & Agung, B. (2020). Asesmen pembelajaran berbasis google form pada mata pelajaran sejarah kebudayaan islam di MAN 2 Bandung. *AL-IBANAH*, *5*(1), 1–27.

Saido, G. M., Siraj, S., Nordin, A. B. Bin, & Al_Amedy, O. S. (2018). Higher order thinking skills among secondary school students in science learning. *MOJES: Malaysian Online Journal of Educational Sciences*, *3*(3), 13–20.

Sanjaya, M. E., Asyhar, R., & Hariyadi, B. (2015). Pengembangan Instrumen Evaluasi pada Praktikum Uji Enzim Katalase di SMA Negeri Titian Teras Muaro Jambi. *Edu-Sains: Jurnal Pendidikan Matematika Dan Ilmu Pengetahuan Alam*, *4*(2).

Sari, D. R. U., Wahyuni, S., & Bachtiar, R. W. (2018). Pengembangan Instrumen Tes Multiple Choice High Order Thinking Padapembelajaran Fisika Berbasis E-Learning Di Sma. *Jurnal Pembelajaran Fisika*, *7*(1), 100–107.

Solichin, M. (2017). Analisis daya beda soal, taraf kesukaran, validitas butir tes, interpretasi hasil tes dan validitas ramalan dalam evaluasi pendidikan. *Dirasat: Jurnal Manajemen Dan Pendidikan Islam*, *2*(2), 192–213.

Sudijono, A. (2001). *Pengantar evaluasi pendidikan*.

Sulianto, J., Cintang, N., & Azizah, M. (2018). *Higher Order Thinking Skills (HOTS) Siswa pada Mata Pelajaran Matematika di Skolah Dasar Pilot Project Kurikulum 2013 di Kota Semarang*.

Suryabrata, S. (2000). Pengembangan alat ukur psikologis. *Yogyakarta: Penerbit Andi*.

Suwarto, S. (2017). Tingkat Kesulitan, Daya Beda, dan Reliabilitas Tes Biologi Kelas 7 Semester Genap. *Seminar Nasional MIPA 2016*.

Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, *10*(11), 78–85.

Thiagarajan, S. (1974). *Instructional development for training teachers of exceptional children: A sourcebook.*

Thomas, A., & Thorne, G. (2009). How to increase higher order thinking. *Metarie, LA: Center for Development and Learning*, 264.

Utomo, D. W. (2015). Pengembangan Sistem Ujian online soal pilihan ganda dengan menggunakan software wondershare quiz creator. *Inovasi Pendidikan Fisika*, *4*(3).

Wahyuningsih, R., Wahyuni, S., & Lesmono, A. D. (2016). Pengembangan Instrumen Self Assessment Berbasis Web untuk Menilai Sikap Ilmiah pada Pembelajaran Fisika di SMA. *Jurnal Pembelajaran Fisika*, *4*(4), 338–343.