BioEdu
Berkala Ilmiah Pendidikan Biologi
Vol. 9 No. 1 Tahun 2020
Hal: 275-283
https://ejournal.unesa.ac.id/index.php/bioedu

CHARACTERISTICS OF MODIFIED MULTIPLE-CHOICE INSTRUMENT TO MEASURE HIGH ORDER THINKING SKILLS FOR ECOSYSTEM SUBJECT

Armita Hildan Haniyah
Biology Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya
Jalan Ketintang Gedung C3 Lt.2, Surabaya 60231
Email: armitahaniyah16030204080@mhs.unesa.ac.id

Muslimin Ibrahim
Biology Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya
Jalan Ketintang Gedung C3 Lt.2, Surabaya 60231
Email: musliminibrahim@unesa.ac.id

Abstract

The demand in the 21st century emphasizes that student shall possess high-order thinking skills. One of the alternative assessments that can be used to practice higher-order thinking skills is modified multiple-choice. Modified multiple-choice consists of two levels; the first level resembles traditional multiple-choice, while the second level is students' reasons in answering the first level, aiming to encourage higher-order thinking skills. The study aims to describe the characteristics of modified multiple-choice for ecosystem subject. The development of the assessment instrument referred to the ADDIE model, which consisted of five stages, including analysis, design, development, implementation, and evaluation. The developed items are then validated by education and material experts to check the validity of items. The items of a multiple-choice assessment instrument are tested to 32 students of 10th graders at SMAN 19 Surabaya. Data analysis technique employed was descriptive analysis technique, including the analysis of validity and test results to determine the reliability value and level of difficulty. The results suggest that the modified multiple-choice assessment instrument was declared valid with a mode of 4, which was included in the "highly valid" category. The reliability value using Cronbach's Alpha formula reached 0.63. Meanwhile, the difficulty index was evenly distributed, with percentages of 20% (easy), 53% (medium), and 27% (difficult). The modified multiple-choice questions for ecosystem subject are declared valid and reliable to measure the students' abilities.

Keywords: item characteristics, modified multiple-choice, validity, reliability, level of difficulty

INTRODUCTION

Education has a critical role in creating flexible, creative, and proactive young generation in facing nation-building challenges in Indonesia in the 21st century whose learning process reflects the four learning objectives (4C), i.e., critical thinking, creativity, communication, and collaboration (Susilo, 2015). Young generations need to be formed to be skilful in resolving problems, making the right decisions, being creative in thinking, and being capable of expressing their ideas effectively (Warsono and Hariyanto, 2012).

The demand in the 21st century emphasizes that students shall possess high-order thinking skill, and it needs to be taught to students. Higher-order thinking skills are thinking skills that include the ability to analyze, evaluate, and create. This definition is in accordance with the revised version of Bloom's Taxonomy which uses the terms of analysis, evaluation, and creation (Anderson & Krathwohl, 2001).

The reason that reinforces the importance of higher-order thinking skill in the 21st-century education (particularly in Indonesia), is to make Indonesian people become accustomed to thinking critically and have a strong work ethic so that they can compete positively at international level (Poedjiadi, 2010). The importance of mastering higher-order thinking skills is also contained in the points of Graduate Competency Standards (SKL) for High School. SKL on the Regulation of the Minister of Education and Culture Number 20 of 2016 concerning Graduate Competency Standards for elementary and secondary education, among others, also states that Senior High School (SMA/MA) graduates must have creative, productive, critical, independent, collaborative, and communicative thinking and acting skills through a scientific approach. For the sake of the implementation of the assigned SKL, assessment instruments should be aligned with the higher-order thinking skills to motivate students in developing their thinking ability and being capable of following the development of knowledge and technology.

In fact, in general, students' higher-order thinking skills in Indonesia are still far below the rank compared with other countries. The survey results of PISA and TIMS prove that Indonesian students are only capable of reaching the second cognitive level out of the six degrees of thinking on tested questions. This finding indicates that the logical and rational thinking skills possessed by Indonesian students are still low so that the Indonesian annual achievement is ranked low among the participating countries (Sani, 2016). The 2018 PISA study showed that Indonesia ranked 70th out of 78 countries participating in PISA (OECD, 2018). In addition, based on the 2015 TIMSS survey, Indonesia ranked at 44th place out of 49 countries with an average score of 397 for the IPA achievements (IEA, 2015).

The improved quality of the National Exam (UN) applied in Indonesia started to implement HOTS-based questions in 2018. Although the HOTS-based questions had only emerged around 10-15% of the total questions to answer. According to Susetyo (2019), the implementation of HOTS questions in the UN affected the UN results. The students' average score decreased. The average score of the National Exam in 2017 amounted to 53.47, while the average score in 2018 reached 51.76. Meanwhile, the average result of the National Exam in 2019 increased, amounting to 53.00. however, the result was still in the lower average compared with the results of the 2017 National Exam (Puspendik, 2018).

Various facts mentioned earlier indicate that the students' thinking skills in Indonesia remain at a low level. As a reflection of the learning results, according to Suprayitno (2019), as the Head of Research and Development Agency of the Ministry of Education and Culture, it is expected that the results of National Examination can become feedback to increase the quality of classroom learning, such as the evaluation of teaching and learning activities, as well as the development of assessment instruments that are capable of training students to attain higher-order thinking skills.

Based on the observation and interview results conducted in schools, the assessment instruments frequently used by the teachers was written test with multiple-choice items. The reason in applying multiple-choice questions is to ease the scoring process and to obtain more objective score. In this type of question, there are only two possibilities; when they answer correctly, they will obtain a score of 1, and when they answer incorrectly, they will obtain a score of 0, but students have a huge possibility that the students choose the answers by chance. Hence, it can be concluded that multiple-choice questions are less effective to measure higher-order thinking skills (Purwanto, 2010).

One of the assessment alternatives that can be employed to practice the higher-order thinking skill is the two-tier multiple-choice questions whose form of questions was developed by Treagust (2006). According to Adodo (2013), a two-tier multiple-choice is a form of questions that is more sophisticated than multiple-choice questions. The first level resembles the traditional multiple-choice questions, commonly associated with the knowledge statements. The second level resembles the forms of conventional multiple-choice questions, aiming to encourage higher-order thinking skills.

According to Chandrasegaran (2007), the use of the multiple-choice instrument is only to assess content knowledge without considering the reason behind the selected answer. As an improvement of this concept, a multiple-choice instrument was developed by including students' responses and alternative conceptions. Students are required to justify their answers by providing reasons. The provision of reason when answering Modified Multiple-Choice items is a sensitive and effective method to assess meaningful learning.

Cullinane (2011) suggested that the provision of reasons on the second level of two-tier multiple-choice questions can improve higher-order thinking skills and observe students' abilities in providing reasons. The provision of reason on the second level of questions can reduce the habit of answering questions by chance that often becomes the weakness of multiple-choice questions in general. The two-tier multiple-choice question provides an objective, easy, and quick scoring compared with other tests for higher-order thinking skill, such as an essay.

The teachers are capable of identifying students' skills well when they employ the appropriate measurement instrument. According to Arikunto (2008), the test requirements to be included as a "good" category must fulfil some criteria, including validity, reliability, objectivity, practical and economical aspects. Out of the five criteria, an assessment instrument can be categorized as "good" when the instrument is, at least, valid and reliable. Validity is the accuracy of the assessment instrument for any matter to assess; therefore, the instrument precisely assesses what to assess (Sudjana, 2010). Validity can be interpreted to the extent of the accuracy or the precision of the measuring instrument in carrying out its measurement function. A valid instrument produces valid data, as well (Widoyoko, 2009).

In addition to validity, reliability is another factor to determine that an instrument is in the "good" category.

Reliability is related to the consistency of test results (Arikunto, 2015). A reliable assessment instrument generates a relatively equal or consistent assessment, even when used repeatedly. The difficulty level of questions on the developed assessment instrument also needs to be identified. It aims to determine whether the questions developed are too easy or too difficult for students. In general, a question with a good category has a medium level of difficulty. According to Widoyoko (2014), the percentage of proper levels of difficulties for test comprised of 25% of difficult questions, 50% of medium questions, and 25% of easy questions, or in other words, the ratio of easy: medium: difficult questions is 1:2:1. Therefore, if an assessment instrument could be answered correctly by all students, it is considered as an easy test with improper questions, and vice versa. (Bagiyono, 2017)

Based on the explanation, this study aims to describe the modified multiple-choice assessment instrument to measure valid and reliable higher-order thinking skills for Ecosystem subject at 10th grade of Senior High School, including the questions' difficulty levels.

METHOD

This research was development research, i.e., developing modified multiple-choice questions to measure students' higher-order thinking skills. The development of the assessment instrument referred to the ADDIE model, which consisted of five stages, including analysis, design, development, implementation, and evaluation. This research was conducted in November 2019 – March 2020. The research object was 15 modified multiple-choice questions that had been validated by experts and declared as valid questions. The assessment instrument trial was conducted to 32 10th graders of SMA Negeri 19 Surabaya who were considered to represent the population to measure the reliability of modified multiple-choice assessment instrument.

The data collection methods used were validation and test methods. The assessment instrument validity was measured based on the results of the validation carried out by education and material experts. Validation activities carried out to determine the validity of the developed assessment instrument. While the test method is carried out by testing the modified multiple choice instrument items on student. From the results of these tests will be known reliability and difficulty level of items developed.

The data analysis technique employed was descriptive analysis technique, including the validity analysis of modified multiple-choice assessment instrument using a Likert scale. The assesment instrument was categorized valid if the average score obtained $\geq$ 3 by the Likert scale. The analysis of modified multiple-choice test results to determine reliability value was carried out by using the Cronbach's alpha formula. The reliability value generated from Cronbach's Alfa formula was then interpreted using the criteria of test reliability level. The developed assesment instrument was categorizes reliabel if the score obtained by $\geq$ 0,60. The value of the difficulty levels was obtained by comparing the number of students who answered correctly divided by the number of students taking the test, then multiplied by 100%. Then the value of the calculation results obtained is interpreted using the item difficulty level criteria according to Arikunto (2015).

The specifications of modified multiple-choice questions to measure higher-order thinking skills for Ecosystem subject is presented in **Table 1**

Table 1. Specifications of modified multiple-choice questions for Ecosystem subject

Basic Competence	Indicators	Assessment Form	Question Number	Cognitive Level
3.10. Analyzing the components of ecosystem and interactions among components	3.10.1 Analyzing the components of an ecosystem, based on images	Two-Tier Multiple-Choice Question	1	C4 (Analyzing)
	3.10.2 Analyzing the component viability of an ecosystem	Two-Tier Multiple-Choice Question	15	C4 (Analyzing)
	3.10.3 Analyzing possibilities that occur on the components of the ecosystem due to imbalances that occur in the ecosystem	Two-Tier Multiple-Choice Question	2	C4 (Analyzing)
	3.10.4 Analyzing the energy flow (food chains and food webs) in an ecosystem	Two-Tier Multiple-Choice Question	3	C4 (Analyzing)
	3.10.5 Analyzing examples of interaction patterns in ecosystems based on data	Two-Tier Multiple-Choice Question	4	C4 (Analyzing)
		Two-Tier Multiple-Choice Question	5	C4 (Analyzing)
	3.10.6 Categorizing the interactions that occur in an ecosystem, based on images	Two-Tier Multiple-Choice Question	10	C6 (Creating)

	3.10.7 Summing up the effects of abundant nitrogen gas in the air towards organisms on earth in the nitrogen cycle	Two-Tier Multiple-Choice Question	13	C5 (Evaluating)
	3.10.8 Summing up the correlation of evapotranspiration and net primary productivity on various types of ecosystems	Two-Tier Multiple-Choice Question	9	C5 (Evaluating)
	3.10.9 Predicting the results of water quality tests that experience the phenomenon of ecosystem imbalance	Two-Tier Multiple-Choice Question	14	C5 (Evaluating)

Basic Competence	Indicators	Assessment Form	Question Number	Cognitive Level
3.10. Analyzing the components of ecosystem and interactions among components	3.10.10 Finding the right reasons to explain the disruption of ecological balance	Two-Tier Multiple-Choice Question	11	C4 (Analyzing)
	3.10.11 Determining attitudes in maintaining the ecological balance	Two-Tier Multiple-Choice Question	6	C5 (Evaluating)
		Two-Tier Multiple-Choice Question	8	C5 (Evaluating)
	3.10.12 Tackling the growth of organisms with negative impacts on the ecosystem	Two-Tier Multiple-Choice Question	7	C6 (Creating)
	3.10.13 Designing solutions for the reduction of environmental carrying capacity	Two-Tier Multiple-Choice Question	12	C6 (Creating)

RESULTS AND DISCUSSION

The questions developed are Modified Multiple-choice type, which consists of two levels. The first level is in the form of stimulus, and each question is accompanied by choices. On the other hand, the second level consists of the reasons underlying the answers chosen at the first level. The stimulus questions developed contain data from research results, images, graphics, and cases that are often encountered in daily life. Thus accustoming students to develop critical and creative thinking skills.

The preparation of the Modified Multiple-choice assessment instrument was conducted by elaborating basic competencies into some indicators and followed with developing the questions by exploring references for the stimulus questions developed from various sources and in accordance with the indicators. Subsequently, the questions were made along with the answer choices that fit the context to let the students explain the reasons for choosing the answer by describing them briefly

The questions that are successfully developed in the Modified Multiple-choice assessment instrument amounted to 15 questions. The examples of Modified Multiple-choice instrument assessment questions developed are presented in **Figure 1.**

Look carefully at the following research results, then choose the right answer!

A student majoring in Biology conducted a study on "The Antagonism Ability of *Pseudomonas sp.* and *Penicillium sp.* against *Cercospora nicotianae* in Vitro". Based on the results of the Antagonism test and zone of inhibition measurement, the results obtained are as follows:

Table 1. Percentage of inhibitory calculation of *C.Nicotianae* (%)

Microbes	Percentage of inhibition (%)			
	36 hours	48 hours	60 hours	72 hours

Microbes	Percentage of inhibition (%)			
	36 hours	**48 hours**	**60 hours**	**72 hours**
*Penicillium sp.*2	33,3	30	42,8	47,05
Pseudomonas sp. 1	0	10	21,4	29,4
Pseudomonas sp. 2	0	0	21,4	23,5

Sources: Putra, Muhammad., Puwantisari, Susiana (2018)

Based on the research results, what are the interactions occurred between *Penicillium sp.* and *Cercospora nicotianae?*

a. Antibiosis
b. Predation
c. Parasitism
d. Comensalism
e. Competion

Describe the reason briefly!

Figure 1. Examples of questions developed in the Modified Multiple-choice assessment instrument

The stimulus used in these questions were research results from Biology-majoring college students. The developed Modified Multiple-choice assessment instrument required the students to think further since they would not only choose one correct answer. Meanwhile, in the reason section, the students were required to provide reasons for the answers chosen at the first level. This is in accordance with Cullinane (2011) who suggests that the inclusion of reason at the second level of two-tier multiple-choice question form can be used to improve the thinking skills and identify the students' capability in providing reasons.

The item instruments that have been compiled cannot be stated directly either well, therefore a review of the item instrument is needed (Rahmani et al, 2015). A good question should be valid and reliable. Validity is the accuracy of the assessment instrument for everything that is assessed so that it actually assesses what should be assessed (Sudjana, 2010). Validity can be obtained from the results of the accuracy and the measurement results by both education and material experts using pre-designed validation instruments. Validity provides an understanding that the evaluation results must be in line or consistent with what has been evaluated (Agustini, 2016). A test is considered invalid when it fails to provide accurate information regarding the attributes it measures (Azwar, 2016).

The validity of the developed modified multiple-choice assessment instrument was obtained from the results of validation by the education and material experts. In compiling the developed modified multiple-choice validity assessment instrument, three aspects must be focused on, i.e., material aspects, construction aspects, and language aspects. The obtained multiple-choice assessment instrument validation resulted in an overall mode of 4. The resulted mode suggested that the modified multiple-choice assessment instrument was valid, while according to the Likert scale, it was very valid. The following data is the result recapitulation of the modified multiple-choice assessment instrument validation presented in **Table 2.**

Table 2 The Recapitulation Results of the Validation Question on the Modified Multiple-choice Assessment Instrument on Ecosystem Subject

No.	Aspects rated	Validator		Category
		V1	V2	
A.	**Subject Matter**			
1	Questions were in accordance with the indicators developed	4	4	Highly valid
2	Questions were in accordance with the truth of the concept	4	4	Highly Valid
3	Questions presented were in accordance with daily life	4	4	Highly valid
4	There was one correct answer	4	4	Highly Valid
Mode				Highly Valid
B.	**Construction**			
1	The instructions were easy to understand	4	4	Highly valid
2	Questions consisted of two levels: 1) answer choices, and 2) inclusion of reasons	4	4	Highly valid
3	Questions employed stimuli that attracted the learners to read	3	3	Valid
4	Questions employed contextual stimuli (picture/graphic/text/corresponding to the real world)	4	4	Highly valid
5	Questions measured the cognitive levels of C4 (analyze), C5 (evaluate), and C6 (create)	4	4	Highly valid
6	Questions did not lead a double interpretation	4	4	Highly valid
7	Questions did not depend on previous questions	4	4	Highly valid
8	Answer choices did not use the statement "all answers are true/false", and the like	4	4	Highly valid
Mode				Highly valid
C.	**Language**			
1	Using Indonesian in accordance with the rules	3	3	Valid
2	Using language that was communicative and easy to understand	4	4	Highly Valid
3	Not containing words/expressions that led to double interpretation/misunderstanding	4	4	Highly valid
4	Not containing ambiguous language	4	4	Highly Valid
5	The answer choices did not repeat the word/group of words unless it was a unity of meaning.	4	4	Highly valid

No.	Aspects rated	Validator		Category
		V1	V2	
Mode				Highly Valid

The first aspect was the material aspect, consisting of four criteria validated by material experts and education experts, i.e., 1) the suitability of the questions with the developed indicators, 2) the relevance of the questions with the truth of the concept, 3) the suitability of the questions with daily life, and 4) there was only one correct answer in each question. This aspect obtained the mode score of **highly valid**. The material aspects are related to science and students' cognitive level (Retnawati, 2016)

The second aspect was the construction aspect, consisting of 8 categories, i.e., 1) easy-to-understand instructions, 2) two-level questions, 3) the stimuli used in questions attracted students to read, 4) the stimuli used were contextual in the form of images/graphics/text/etc., 5) questions measured cognitive level of C4 (analyze), C5 (evaluate), and C6 (create), 6) questions did not lead to multiple interpretations, 7) questions did not depend on previous questions, 8) answer choices did not use the statement "all answers are true/false" and the likes. The construction aspect was related to the technique of writing questions (Mardapi, 2017). This aspect obtained the total score with the category of **highly valid**. However, in point no. 3 (questions employ stimuli that attract the learners to read), there were several numbers of questions that were less appropriate, i.e., question number 1, 11, and 13.

The third aspect was the aspect of language consisting of 5 categories, including 1) using proper Indonesian, 2) using communicative and easy-to-understand language, 3) not containing words/expressions that led to double interpretation, 4) not containing ambiguous language, 5) the answer choices did not repeat the word/group of words unless it was a unity of meaning. This aspect obtained a total score mode in the category of **highly valid**. However, in the point of using proper Indonesian, there were several numbers of questions that were not entirely appropriate, i.e., questions number 2, 4, 5, 6, and 10. The language aspects were associated with the clarity of every aspect supporting the question preparation (Mardapi, 2017). The assessment instrument which was theoretically stated valid indicated that it had fulfilled all three aspects (Rachma and Ratnasari, 2015).

After being tested for its validity, the instrument development of 15 modified questions was then examined in a limited manner to be analyzed later. The analysis included the questions' reliability and difficulty levels. Reliability is related to the determination or the severity of the test results (Arikunto, 2015). The data obtained from the test results were utilized to determine the reliability value using the Cronbach Alpha formula. The reliability value was then interpreted using the test reliability level criteria. The calculation result of the reliability value using the Alfa Cronbach formula resulted in the score 0.63 with the high category. Thus, the resulting multiple-choice instrument can be considered as valid and reliable. A measuring instrument is declared to have a high coefficient of reliability if it provides an equal or almost equal value when it is used to assess the same object at different times.

The questions' reliability is related to the consistency of the questions to measure students' learning abilities (Masruroh, 2012). Several factors affect the reliability value, both directly and indirectly. The direct factors include the test implementation time, the questions' difficulty levels, the questions' length, the scoring objectivity, and the answer and score dissemination, whereas the indirect factors include clear implementation instructions and environmental conditions and supervision (Retnawati, 2016).

The question difficulty index can be employed as one of the parameters for analyzing a test since it has a function to determine the student's ability (Retnawati, 2016). The difficulty level calculation aims to discover whether the questions are too difficult or too easy for students through calculations by comparing the number of students who answered correctly with the total number of students who took the test. The questions difficulty level are determined by the number of participants who answered correctly divided by the number of students who took the test, then multiplied by 100% (Maenani and Oktova, 2015). The following table and diagram explain the recapitulation results of the modified questions difficulty level

Table 3. The Difficulty Level Distribution of Multiple-Choice Questions Item on Ecosystem Subject Modification

Question No.	Difficulty level	Criteria
1	0.25	Difficult
2	0.25	Difficult
3	0.28	Difficult
4	0.40	Medium
5	0.71	Easy
6	1.16	Easy
7	0.63	Medium
8	0.66	Medium
9	0.28	Difficult
10	0.41	Medium
11	1.16	Easy
12	0.44	Medium
13	0.47	Medium

14	0.56	Medium
15	0.69	Medium

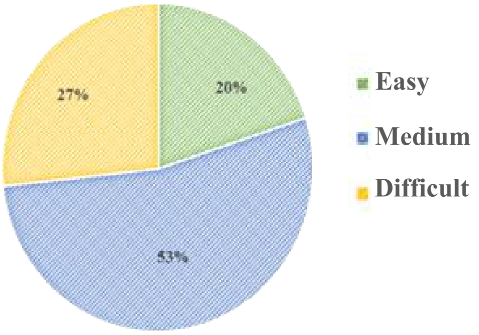
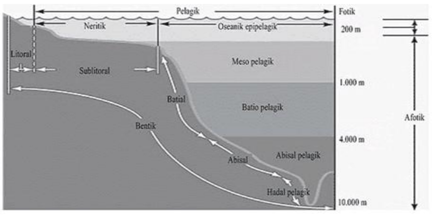


Figure 2. Difficulty Level Distribution of Multiple-Choice Questions Item on Ecosystem Subject Modification

The calculation of the difficulty level functions to determine whether the questions are too difficult or too easy for students. The calculation was made by comparing the number of students who answered correctly with the total number of students who took the test. The difficulty level of the questions item that has been obtained through subsequent calculations is interpreted using the difficulty level questions criteria developed by Arikunto (2015). The percentage of difficulty levels on the multiple-choice assessment instruments developed were 20% easy, 53% medium, and 27% difficult. According to Widoyoko (2014), a test given to students should have a distribution balance between easy, medium, and difficult questions with a ratio of 25%, 50%, and 25%. The following figure is the sample questions on multiple-choice assessment instruments that are categorized into difficult criteria. These questions are presented in **Figure 3.**

1. Look at the following picture, then choose the right answer!



Biophysically, the sea area can be divided according to vertical and horizontal dimensions, physical factors, and the distribution of biota communities. Each zone has unique physical, chemical, and biological characteristics. In addition, ocean zoning can be divided into surface zone (pelagic zone) and bottom zone (benthic zone). Horizontally, the pelagic zone is divided into two zones, i.e., the neritic zone and the oceanic zone. Meanwhile, vertically, it is divided into photic zone and aphotic zone. The photic zone is also called as epipelagic zone, while the aphotic zone is divided into four zones, i.e., mesopelagic, batipelagic, abyssal and hadal pelagic zones.

Based on these zones, in which part of the zone are the most abundant producer-level components found?

a. Photic Zone
b. Neritic Zone
c. Aphotic Zone
d. Oceanic Zone
e. Benthic Zone

Describe your reasons briefly!

Figure 3. One of the items of modified multiple-choice that have a high difficulty level

The questions listed in **Figure 3** could only be answered correctly by 8 out of the 32 students taking the test.Thus, the questions obtained a difficulty level of 0.25 with difficult interpretation. The difficulty level value is obtained by comparing the number of students who answered correctly with the total number of students taking the test. Therefore, if the difficulty index is high, the interpretation of the questions are classified into the "easy" category. In contrast, if the difficulty index obtained is low and close to zero, the questions are categorized as "difficult" for students.

CONCLUSION

Based on the research results on the instrument development of the modified multiple-choice assessment to measure students' higher-order thinking skills on Ecosystem subject, it can be concluded that the modified multiple-choice assessment instrument is declared valid and reliable. Validity is obtained based on the results of validation by material experts and education experts by considering material aspects, construction aspects, and language aspects which obtains a mode of 4, which is included in the "highly valid" category. The reliability value is obtained by analyzing student test results using Cronbach's Alpha formula, reaching 0.63. Meanwhile, the difficulty index obtained is heterogeneous, i.e., 20% (easy), 53% (medium), and 27% (difficult). Based on the conducted research, it can be concluded that the items of the modified multiple-choice assessment instrument are valid and reliable.

ACKNOWLEDGMENT

The author would like to thank those who have helped complete this research namely, Prof. Dr. H. Muslimin Ibrahim, M.Pd., Dr. Raharjo, M.Si., Dr. Sunu Kuntjoro, M.Si. for their suggestions and input to realize the development of modified multiple choice instrument to measure high order thinking skills for ecosystem subject.

REFERENCES

Adodo, S. O. 2013. "Effects of Two-Tier Multiple Choice Diagnostic Assessment items on Students' Learning Outcome in Basic Science Technology, Ondo State". *Academic Journal of Interdisciplinary Studies by MCSER-CEMAS-Sapienza University of Rome*. E-ISSN 2281-4612. ISSN 2281-3993, Vol. 2 No. 2

Agustini, R., Nasrudin, H., Azizah, U., dan Muchlis. 2016. *Asesmen.* Yogyakarta : Absolute Media.

Anderson, L.W., dan Krathwohl, D.R. 2001. *A Taxonomy for Learning, Teaching, and Assesing: A Revision of Bloom's Taxonom y of Educatioanl Objectives*. New York: Addison Wesley Longman, In

Arikunto, S. 2008. *Prosedur Penelitian Suatu Pendekatan Praktik.* Jakarta : Rineka Karya

Arikunto, Suharsimi. 2015. *Dasar-dasar Evaluasi Pendidikan Edisi 2*. Jakarta: PT Bumi Aksara

Azwar, S. 2016. *Konstruksi Tes Kemampuan Kognitif.* Yogyakarta : Pustaka Belajar

Bagiyono. 2017. "Analisis Tingkat Kesukaran dan Daya Pembeda Butir Soal Ujian Pelatihan Radiografi Tingkat I". *Jurnal Widyanuklida Vol.16 No.1.*

Beyrak, B. K.. 2013. "Using Two-Tier to Identify Primary Students' Conceptual Understanding and Alternative Conception in Acid Base". *Mevlana International Journal of Education*. Vol. 3, No. 2, pp. 19-26

Chandrasegaran, A. L., Treagus, D. F. Mucerino, M.. 2007. "The Development of a Two-tier Multiple-Choice Diagnostic Instrument for Evaluating Secondary School Students' Ability to Describe and Explain Chemical Reactions Using Multiple Level of Representation". *The Royal Society of Chemistry, Chemistry Education Research and Practice*, Vol. 8, No. 3, 293-307

Cullinane, Alison, dan Maeve, Liston. 2011. *Two-tier Multiple Choice Question: An Alternative Method of Formatif Assessment for First Year Undergraduate Biology Students*. Limerick: NationalCenter for Excellence In Mathematicsand Education Science Teaching and Learning (NCE-MSTL)

IEA. 2015. *Student Achievement Overview (Science) Grade 4.* http://timss2015.org/timss2015/science/student-achievement

Maenani, L. dan Oktova, R. 2015. "Analisis Butir Soal Fisika Ulangan Umum Kenaikan Kelas X". *Jurnal Berkala Fisika Indonesia Vol. 7 No.1.*

Mardapi, D. 2017. *Pengukuran, Penilaian, dan Evaluasi Pendidikan (edisi 2)*. Yogyakarta: Parama Publishing.

Masruroh., Rudyatmi, Ely., dan Ridlo, Saiful. 2012. Analisis Soal Ulangan Semester Gasal Biologi Kelas X di Kecamatan Petanahan Kebumen. *Unnes Journal of Biology Education*. Vol. 1 (2); 116-121

OECD. 2018. *PISA 2018 : PISA Result In Focus*. Diakses pada 2 Oktober 2019. https://www.oecd.org/pisa.

Permendiknas No 20. 2016. *Standar Kompetensi Lulusan untuk Satuan Pendidikan Dasar dan Menengah.* Jakarta: Depdiknas

Poedjiadi, Anna. 2010. *Sains Teknologi Masyarakat: Model Pembelajaran Kontekstual Bermuatan Nilai*. Bandung: PT Remaja Rosdakarya bekerjasama

Purwanto, Ngalim. 2010. *Prinsip- prinsip dan Teknik Evaluasi Pengajaran.* Bandung: PT Remaja Rosdakarya

Puspendik. 2018. *Laporan Hasil Ujian Nasional*. Diakses dari https://puspendik.kemdikbud.go.id/hasil-un/ (9 November 2019)

Rachma, N. A dan Ratnasari, Evie. 2015. Pengembangan Tes Elektronik (E-Test) Berbasis Komputer pada Materi Bioteknologi di SMA Negeri 1 Surabaya. *Jurnal Bioedu Unesa*. Vol. 4 (3). page 1018-1022

Rahmani, Mita. 2015. Analisis Kualitas Butir Soal Buatan Guru Biologi Kelas X SMA Negeri 1 Tanah

Pinoh. *Artikel Penelitian Pendidikan*. Pontianak: Universitas Tanjungpura.

Retnawati, Heri. 2016. *Validitas, Reliabilitas, dan Karakteristik Butir (Panduan untuk Peneliti, Mahasiswa, dan Psikometrian)*. Yogjakarta: Parama Publishing.

Rositasari, D., Saridewi, N. & Agung, S., 2014. "Pengembangan Tes Diagnostik Two-Tier untuk Mendeteksi Miskonsepsi Siswa SMA Pada Topik AsamBasa". *Edusains*, Volume 6, pp. 169-176.

Sani, R. A. 2016. *Penilaian Autentik*. Jakarta: Bumi Aksara

Sudjana, N. 2010. *Penilaian Hasil Proses Belajar Mengajar*. Bandung: PT Remaja Rosdakarya

Suprayitno, Totok. 2019. Kemendikbud: Nilai UN Tahun ini Mengalami Kenaikan. Diakses pada tanggal 30 Januri 2019. https://news.okezone.com/read/2019/05/08/65/2052992/kemendikbud-nilai-un-tahun-ini-mengalami-kenaikan

Susetyo, Bambang Agus. 2019. Soal HOTS Tetap Mewarnai UN 2019. Diakses pada 2 Oktober 2019.https://lpmpjatim.kemdikbud.go.id/.

Susilo. 2015. "Curriculum of EFL Teacher Education and Indonesian Qualification Framework: A Blip of the Future Direction". *Jurnal Dinamika Ilmu*, 15 (1): 11-24

Treagust, David F. 2006. *Diagnostic Assesment In Science as A Means to Improving Teaching, Learning, and Retention*. UniServe Science Assesment Symposium Proceedings. The Universityof Sydney

Tuysuz, C. 2009. "Development of Two-tier Diagnostic Instrument and Assess Students Understanding in Chemistry". *Academic Journal of Scientific Research and essay*, 4(6) ISSN 1992- 1248, 626-631.

Warsono dan Hariyanto. 2012. *Pembelajaran Aktif.* Bandung: PT. Remaja Rosdakarya.

Widoyoko, S.E.P. 2009. *Evaluasi Program Pembelajaran (Panduan praktis Bagi Pendidik dan Calon Pendidik).* Yogyakarta : Pustaka Pelajar.

Widoyoko, S.E.P. 2014. *Penilaian Hasil Pembelajaran di Sekolah.* Yogyakarta : Pustaka Belajar