

Two-Tier Test: Development of Critical Thinking Instruments for Ecosystem Concepts

Helda Dumayanti¹, Risda Putri Indriani², Evita Nury Hariyanti³, Rizhal Hendi Ristanto⁴, Mieke Miarsyah⁵

^{1,2,3,4,5}Pendidikan Biologi, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Jakarta, Indonesia

Abstract

Critical thinking is necessary in improving the quality of education. Students' critical thinking includes interpretation, analysis, evaluation, inference, interpretation, and self-regulation. The two-tier test instrument was developed to determine students critical thinking skills in the ecosystem at class X MIPA di MAN 1 Bogor City. This research method used was based on 4D development model (Define, Design, Develop, and Disseminate). The instrument was tested on two expert validators and 35 students. To test the success rate of the test, validation, reliability, and item analysis were used. The item analysis used is the level of difficulty of the questions and differentiating power of answer choice. Based on the results of the product trial, 10 valid questions and 5 unvalid questions. The reliability test result was 0,671 which the interpretation that 45% of respondents considered the instrument developed to be reliable. In addition, 10 questions are in the difficult category and 5 questions are in the medium category. The differentiating power of answer choice results was 1 question in the very good category, 6 questions in the good category, 4 questions in the moderate category, and 4 questions in the poor category. The two-tier test instrument for ecosystem critical thinking in this study is declared valid and reliable so this can be used as an instrument in student's learning.

Kata kunci : 21st century skills, assessment, critical thinking

Two-tier Test : Pengembangan Instrumen Berpikir Kritis Konsep Ekosistem

Abstrak

Berpikir kritis diperlukan dalam meningkatkan kualitas pendidikan. Berpikir kritis siswa meliputi interpretasi, analisis, evaluasi, kesimpulan, interpretasi dan regulasi diri. Instrumen *two tier test* ini dikembangkan untuk mengetahui keterampilan berpikir kritis ekosistem siswa kelas X MIPA di MAN 1 Kota Bogor. Penelitian ini menggunakan model pengembangan 4D (*Define, Design, Develop, and Disseminate*). Instrumen diujikan kepada dua validator ahli dan 35 siswa. Untuk menguji tingkat keberhasilan tes, maka digunakan uji validasi, reliabilitas, dan analisis butir soal. Analisis butir soal yang digunakan adalah tingkat kesulitan soal dan daya beda. Berdasarkan hasil uji coba produk diperoleh 10 soal yang valid dan 5 soal tidak valid. Hasil reliabilitas tes sebesar 0,671 dengan interpretasi bahwa 45% responden menilai instrumen yang dikembangkan dapat dipercaya (reliabel). Selain

itu, 10 soal dengan kategori sulit, dan 5 soal kategori sedang. Daya beda soal menunjukkan 1 soal kategori baik sekali, 6 soal kategori baik, 4 soal kategori cukup, dan 4 soal kategori kurang. Instrumen *two tier test* berpikir kritis ekosistem dalam bentuk pilihan ganda yang telah dikembangkan pada penelitian dinyatakan valid dan reliabel sehingga dapat digunakan sebagai instrumen dalam pembelajaran siswa.

Kata kunci: Evaluasi, Berpikir Kritis, Keterampilan Abad 21

INTRODUCTION

Evaluation of e-learning-based biology learning can assist teachers in designing assessment systems and checking student learning outcomes (Zahara, 2015). Learning evaluation is directed at the components of the input, process and output of learning with a variety of student growth assessment processes. Learning must pay attention to the principles, benefits, requirements and objectives of conducting evaluations (Magdalena et al., 2020). The purpose of the evaluation is to determine the effectiveness and efficiency of learning (Ngafifah, 2020).

21st century learning prepares students to grasp deep knowledge and apply effective critical thinking skills to tackle challenges in an ever-changing society. Critical thinking must be mastered before problem solving, creative thinking, and decision making are carried out (Ikhsanudin & Subali, 2018). Teachers need to develop critical thinking in the learning process and make evaluations of biology learning that can open students' mindsets. Students' critical thinking skills can be trained through providing practice questions that require critical thinking efforts (Aripin, 2017).

Many researchers define critical thinking, among others, as analyzing arguments or evidence (Facione, 2015) solving a problem or making a decision (Duchscher, 1999) or as a process in which a person questions all aspects of a situation and is critical of it (Simpson & Courtney, 1999). 2002). Based on the research results, students' average critical thinking skills are low because they have not reached the minimum standard of completeness (Allanta & Puspita, 2021; Harahap et al., 2020).

According to Facione, students' critical thinking includes interpretation, analysis, evaluation, inference, interpretation and self-regulation. Critical thinking can explain what is thought and how to achieve that assessment. The ideal critical thinker is curious, broad-minded, open-minded, flexible, fair judgment, wise judgment, willing to rethink, eager to find relevant information and persistent in seeking research results (Facione,

2015). The results of developing critical thinking skills from research on biological critical thinking in class XI will increase students' ability to access information and define problems based on accurate facts and data (Sari & Paidi, 2019).

Critical thinking is important to improve the quality of education. Currently, quality education focuses more on the extent to which students' performance on teacher-made tests can predict their potential performance on standardized tests. This test as a test for formative assessment and evaluation. In recent years, assessment is a central issue in the field of education and is often discussed by many stakeholders from the classroom level at school, regional, national and international levels. Educational assessment is important to do to obtain data on the extent to which the level of achievement of educational goals is implemented (Ikhsanudin & Subali, 2018). The existence of a two tier test is expected to help teachers to prepare learning according to the abilities of students (Eriza et al., 2020). Research on the two-tier test was also conducted to assess students' conceptual understanding of ecosystem material (Firman et al., 2021).

At the school level, formal learning assessment refers to a curriculum designed in the form of learning for students (Ikhsanudin & Subali, 2018). The importance of acquiring higher order thinking skills can help students apply information, make decisions, analyze and solve complex problems (Noviana et al., 2014). Capturing inference as a sub-question is used to improve higher order thinking skills and identify students' critical thinking skills (Hala et al., 2019). The purpose of this second level is to encourage students to acquire high-level thinking and reasoning skills (Yamtinah et al., 2016). One of the keys to the success or failure of a learning can be seen from the evaluation or assessment. With the evaluation, we can find out to what extent students' understanding, as input for teachers regarding the learning process, but can also be used as a value.

Most forms of evaluation questions are multiple choice. Its easy nature in making and processing and the results of student answers can come out quickly, making teachers often use multiple choice models both on tests and exams (Dibattista & Kurzawa, 2011). In multiple choice questions, students often guess the answer so that it is not valid to determine the level of students' critical thinking. Treagust developed a two-tier or two-tier assessment, in which the first-level and second-level multiple choice questions select reasons for the first-level questions (Treagust, 1986). Two-tier assessment has been used

in various disciplines including science. This is in line with st udi previous research that has been done on the subject of Biology.

Various two-tier assessment studies have been carried out including on archaea and bacteria material (Kurniasih & Haka, 2017), cell reproduction and division (Sesli & Kara, 2012), genetics (Tsui & Treagust, 2010), photosynthesis (Griffard & Wandersee, 2001). ; Haslam & Treagust, 1987; Treagust, 1986), global warming (Suryawirawati et al., 2018), and environmental changes (Haka et al., 2019). Critical thinking is very important as a tool of inquiry and a pervasive and self-correcting human phenomenon (Facione, 2015). The learning process has not been implemented optimally in helping students to have critical thinking skills (Husna et al., 2021). Rukayah's research results show that the two-tier test is effective in determining concept understanding and is an alternative for assessing student learning outcomes (Rukayah et al., 2020).

The instrument that will be developed in this research is a two tier test instrument in the form of multiple choice to measure ecosystem critical thinking skills. The questions in the assessment will refer to critical thinking indicators, which include interpretation, analysis, interference, explanation and self-regulation (Facione, 2015). Based on this statement, it is necessary to conduct research on "Two-tier Test: Development of Critical Thinking Instruments for Ecosystem Concepts"

RESEARCH METHODS

The research method uses R&D with the Thiagarajan 4D development model, but this research does not use the disseminate stage.

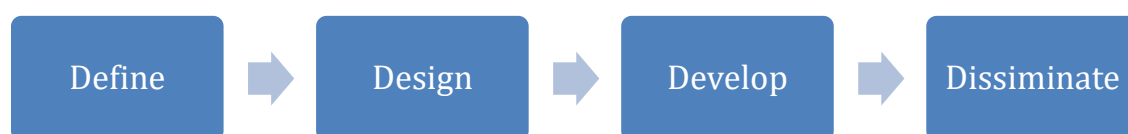


Figure 1. 4D Development Model Chart

The define stage aims to collect information related to the development of the instrument and analyze it such as analyzing core competencies, basic competencies, and making question grids based on material sub-concepts and indicators of critical thinking skills. The design stage is the planning or design stage to determine ecosystem sub-concepts, measured evaluation indicators and critical thinking, formulation of level 1 questions,

formulation of level 2 questions, answer keys, discussion and reference sources used. In addition, it also creates a product assessment sheet for expert validation to make it easier for the assessment.

Table 1. Indicators of critical thinking skills according to Facione in ecosystems

Interpretation	Interpret and classify ecosystem components, interactions between components and various ecosystems
Analysis	Analyzing the interactions between ecosystem components, the ecological pyramid
Conclusion	Draw conclusions about food chains, food webs and ecological pyramids
Evaluation	Evaluating the components that make up the ecosystem and various ecosystems
Explanation	Explain about food chains, food webs, biogeochemical cycles
Self Regulation	Estimating the impact on problems related to biogeochemical cycles

The develop stage is by developing the instrument into two levels of test questions (Two-tier test) in accordance with critical thinking indicators. To determine the success of the test in measuring students' critical thinking level, validity, reliability, and item analysis were conducted.

Expert validation is carried out by biology education lecturers and certified biology teachers using the validity rubric. The validation rubric consists of assessments, namely material, construction and language. The analysis uses a scale of 1-4. Instruments that have been tested by expert validators and declared valid, can be tested on students to get empirical validation. The trial was carried out on 35 students of class X MIPA at MAN 1 Bogor City 2021/2022. The trial technique was carried out using random sampling, which was carried out randomly regardless of class strata. Scoring is done on a scale of 0-1. Students who gave wrong answers at one level or both levels were given a score of 0. Students who gave correct answers at both levels were given a score of 1. The results of empirical validation were then processed by the biserial point correlation validation test formula. A test question is said to be valid if $r_{count} > r_{table}$ ($\alpha = 0.05$ or $= 0.01$). Furthermore, reliability testing is carried out. Data that gives consistent results can be said to be reliable. A test is considered reliable if the reliability is high and the standard error of measurement is small. Reliability test using the KR-20 formula.

The item analysis was carried out by calculating the level of difficulty of the questions and the power of difference. The difficulty level of the questions is the ability of students to answer questions correctly at a certain level of proficiency. The goal is to

provide an introduction to concepts that require re-teaching. Meanwhile, the discriminatory power of questions is the ability of the items to be able to distinguish between students who have mastered the material and students who have not mastered the material. The higher the discriminatory index, the more likely the question is to distinguish between students who understand and do not know.

RESULTS AND DISCUSSION

The instrument was developed with the Thiagarajan 4D development model. In the define stage, information about the test instrument is based on the curriculum and syllabus for class 10 SMA. Ecosystem material is selected based on the basic competencies contained in the 2013 curriculum, namely 3.5 Analyzing ecosystem components and interactions between these components and 3.6 Analyzing data on environmental changes, their causes, and their impacts on life

Table 2. Written Test Grid

Sub-Concept	Critical Thinking Indicator						Σ Questions
	1	2	3	4	5	6	
Components of the ecosystem	1			1			2
Interaction between components	1	1			1		3
Food chains and food webs			1		1		2
Ecological pyramid		2	1				3
Kinds of ecosystem	1			1			2
Biogeochemical cycle					1	2	3
Total Questions	3	3	2	2	3	2	15

Description: (1) Interpretation, (2) Analysis, (3), Conclusion, (4) Evaluation (5) Explanation, (6) Self Regulation (Facione, 2015).

The design stage begins with determining sub-materials, question indicators, critical thinking indicators, and formulating level 1 and 2 questions, making answer keys and discussion. Designing a two tier test instrument for critical thinking skills in the form of multiple choice. The critical thinking skill instrument developed consisted of 15 multiple choice questions. All items of the instrument have been tested for validity and reliability so that they can represent all indicators of critical thinking skills.

Table 3. Instruments and Indicators of Critical Thinking Skills

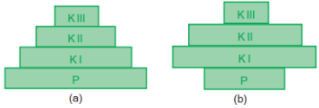
<p>Sub Material: Various ecosystems</p>	<p>Question Indicator : Students can explain the meaning and animals that live in the tundra ecosystem</p>	<p>Key : C-D Critical Thinking Indicator : Interpretation</p>
<p>Formulation of level 1 questions The following include examples of animals that are in the tundra ecosystem are a. cow, deer, giraffe b.snake, crocodile, fish c.penguin, bear, wolf d.tiger, tiger, elephant e.shark, whale, squid</p>	<p>Formulation of level 2 questions Reason: a.an ecosystem that is mostly filled with water b.Ecosystems with high rainfall c.hot and arid ecosystem d.low temperature ecosystem e.an ecosystem where only one type of plant grows</p>	<p>Discussion : Tundra is an ecosystem that has low rainfall and a low temperature of less than 10°C. Animals that live in the tundra ecosystem are penguins, bears and wolves. Source : Campbell, Neil A 2008.. <i>Biologi Jilid 3.</i> Jakarta: Erlangga.</p>
<p>Sub Material: Ecological pyramid</p>	<p>Question Indicators: Students can analyze the comparison of 2 paramida</p>	<p>Key : A - E Critical Thinking Indicator : Analysis</p>
<p>Formulation of level 1 questions</p>  <p>The pyramid of numbers describes the size of the population density in an area. Of these pyramids, which pyramid is the most balanced?</p> <p>a. pyramid (a), because there are more producers than consumers 1 b. pyramid (a), because consumer 1 can eat consumer 2 c. pyramid (b), because consumer needs 1 can be met d. pyramid (b), because consumer 2 has an abundance of food e. pyramid (b), because food for consumers is still fulfilled</p>	<p>Formulation of level 2 questions Reason: a. consumer 2 dominates so that it becomes enough food for other consumers. b.Producers that are few in number have no effect on the balance of the pyramid because they are easy to breed c.pyramid (b) will be profitable for consumers 3 d.consumer 3 has an abundance of food in the pyramid e.producers act as providers of food, so there must be plenty of them</p>	<p>Discussion : The most balanced pyramid is pyramid A, because there are more producers than 1 consumers and so on. Producers act as providers of food for other organisms. Pyramid B is unbalanced because there are 1 more consumers than producers. So that if the producer will run out if there are too many consumers Sumber : Anshori, Moch., dll. 2009. <i>Biologi 1: Untuk siswa menengah atas-madrasah Aliyah.</i> Jakarta: Pusat Perbukuan.</p>

Table 3 shows an example of a two tier test instrument to measure students' critical thinking skills. Two tier test has two levels of multiple choice, the first level asks students' critical thinking knowledge about concepts. The second level is students' reasoning from the answering process at the first level (Sudirman et al., 2021). The second level is reasoning about the reasons for choosing the answer at the first level. The two tier test instrument helps in testing students' higher thinking levels than with ordinary multiple choice questions (Laksono, 2018). The product development or development stage produces a two-tier test instrument to measure critical thinking skills on the ecosystem concept. To determine the success of the product, validity, reliability and item analysis were conducted. The instruments tested were 15 two-level questions.

The validation test was carried out in two steps, namely expert validation and empirical validation. Expert validation is carried out by two experts. The first validator is an education lecturer who is an expert in biology and ecosystem education. The second validator is by a certified biology education teacher. Expert validation is done by comparing the contents of the instrument, indicators of ecosystem questions, and indicators of critical thinking skills.

Table 4. Suggestions from expert validators

No.	Expert Validator Suggestions
1.	There is some ambiguous use of language
2.	Inappropriate answer choices
3.	The picture is not clear
4.	More variety in making questions

The development of this instrument is also assessed by the validator. Aspects assessed by the validator are material, construction and language aspects. The average results of the two validators then become a benchmark for further to the next stage, presented in table 5 below:

Table 5. Expert Validation Results

Rated aspect	Percentage	Criteria
Materi		
- Relevance between material sub-concepts and critical thinking indicators	95%	Very valid
- Formulation of questions and answer keys		
- Relevance between material sub-concepts and measurement		
- Relevance between sub-concepts with level, type of school, and grade level		

Rated aspect	Percentage	Criteria
- Use of the right words and language		
Construction		
- Clarity of questions and answer choices	92,5%	Very worth it
- Instructions from the questions given with the correct answer choices		
- Homogeneous answer choices		
- There is only one correct answer key		
- Case descriptions, tables, figures or graphs work well		
Language		
- Communicative questions	97,5%	Very worth it
- Sentences using standard spelling according to EYD and KBBI		
- Sentences have no double meaning and misperception		
- Using a familiar language (not a local language or a new absorption language that is not well known)		
- Sentence does not offend the respondent		
Average	95%	Very worth it

From table 5, information is obtained that expert validation produces a two-tier test instrument to measure critical thinking skills in very feasible criteria and there are several suggestions to improve the instrument. After repairs are made, empirical validation is carried out. Empirical validity is carried out to find similarities between research results and facts in the field. How to collect data using google form and calculated with biserial points. The validation results show 10 valid questions and 5 invalid questions. Questions can be said to be valid if the value of rcount is greater than the value of rtable. The rtable value is 0.329 with an alpha value of 0.05. The results of the 10 valid questions have met the 6 indicators of critical thinking skills that are being tested. This shows that the 10 questions have been able to test students' critical thinking skills. Meanwhile, 5 invalid questions must be corrected to meet the standards in measuring students' critical thinking skills.

The reliability of the test was carried out using the formula of Kuder Richardson-20 (KR-20). The reliability calculation shows a reliability coefficient of 0.671. Based on this, it can be interpreted that 45% of respondents considered the instrument developed to be reliable (reliable), presented in table 6 below:

Table 6. Empirical Validity Results

Criteria	Critical Thinking Indicator	Question Number	Total
Valid	Interpretation	11	1
	Analysis	4, 8, 9	3

Criteria	Critical Thinking Indicator	Question Number	Total
	Conclusion	7, 10	2
	Evaluation	12	1
	Explanation	5, 13	2
	Self Regulation	15	1
Amount			10
Invalid	Interpretation	1, 3	2
	Evaluasi	2	1
	Explanation	6	1
	Self Regulation	14	1
Amount			5

Based on table 6 the results of empirical validation, it is known that 10 of the 15 questions can be declared valid. Validity value based on biserial points. This calculation produces a coefficient value that shows the strength of an item score to support all items (Zein et al., 2013). The 10 questions have exceeded the correlation coefficient value by 0.329 at an alpha of 0.05. This can be interpreted if the problem is applied to another population, it will show the same results. In addition, the resulting reliability coefficient is 0.671 with the interpretation that 45% of respondents consider the developed instrument to be reliable (reliable). Analysis of the items that are calculated is the level of difficulty of the questions and the differentiating power of the questions. Item analysis was carried out to reveal the cause of the invalidity of the item. The level of difficulty of the questions is done to assess the ability of questions that are too easy or too difficult. Questions that are too easy or too difficult will affect the distribution of scores, causing a decrease in the value of the validity of the test to measure critical thinking skills, presented in table 7 below:

Table 7. Result of Problem Difficulty

Criteria	Question Number	Amount
Sedang	4, 5, 6, 7, 8, 9, 10, 13, 14, 15	10
Mudah	1, 2, 3, 11, 12	5
Total		15 Questions

Based on the information in table 7, information about the questions in the easy category indicates that students can work on the questions. The questions in the medium category indicate that the upper and lower groups can answer the questions with the appropriate proportions. That is, students from the upper class tend to answer more than students from the lower class. The index of difficulty level is calculated on each question. In principle, the average value obtained by students on the items is called the item difficulty level. The

function of the level of difficulty of the items for the purpose of the test, namely questions with a moderate level of difficulty and easy to use for examination purposes and measuring students' critical thinking levels.

The discriminatory power index is the ability of items to distinguish between students who already have high and low critical thinking skills. Basically, evaluation activities are carried out to measure students' abilities individually. The results show that the discriminatory power index varies from less, enough, good, and very good. The higher the discriminatory power index, the better the question is able to distinguish students who understand and do not understand the material, as shown in table 8 below:

Table 8. Results of Differential Power Questions

Criteria	Question Number	Amount
Very good	10	1
Good	4, 8, 9, 11, 12, 13	6
Enough	5, 6, 7, 15	4
Not enough	1, 2, 3, 14	4
Total		15 Question

From table 8, the following information can be described; Item analysis is used to uncover the causes of invalid questions. Items that are not valid are items 1, 2, 3, 6, and 14. Items 1, 2 and 3 have a difficulty level of easy questions and a lack of differentiating power index. The questions are easy because the questions being tested are basic concepts that have been mastered by students. So that the question has not been able to distinguish students who already have good critical thinking skills or not. These questions can be discarded and cannot be used for further tests.

Item number 6 has a moderate level of difficulty and sufficient discriminating power. While item number 14 has a moderate level of difficulty and less discriminating power. Although the level of difficulty of the questions is moderate and the power of difference is sufficient, the validity of the questions remains low. This shows that there are other factors that affect the validity of the questions in addition to the level of difficulty of the questions and discriminating power. Other factors such as the use of ambiguous language so that students are confused about choosing the right answer, the time allocated to students is too loose because it is done online and internal factors of students such as fatigue, illness, or other psychological factors. So that questions number 6 and 14 are discarded and cannot be used for further tests. Based on the discussion, it is known that there are 10 questions worthy of being used as an instrument for evaluating students'

critical thinking skills. This instrument was developed so that the evaluation of critical thinking skills can be measured specifically. Students who have critical thinking skills are able to distinguish between facts and opinions that develop in society, are able to make an action solution and analyze problems, as well as develop skills, expand thinking processes and increase concentration (Chukwuyenum, 2013; Rizal, 2017).

It was found that the use of a two-level test evaluation instrument to measure students' critical thinking skills could be done with multiple choice questions. This is in line with Dharmawati's research (Dharmawati et al., 2016; Walid et al., 2021) which developed multiple choice questions for critical thinking evaluation with a one-level test for junior high school students. This study was developed to test the critical thinking skills of high school students. Therefore, the development of this instrument can raise the level of evaluation to a higher stage.

CONCLUSION

Based on the research results, the developed two tier test instrument can be used to determine students' critical thinking skills regarding ecosystem materials. Based on the results of the product trial, 10 valid questions were obtained and the reliability of the test was 0.671 with the interpretation that 45% of respondents considered the instrument developed to be reliable (reliable). Thus, the two tier test instrument for ecosystem critical thinking in the form of multiple choice that has been developed is valid, feasible and reliable. In connection with the results of the development of the instrument, it is hoped that teachers always try to design variations of the two tier test of ecosystem critical thinking instruments at the level of difficulty of easy, medium and difficult questions.

REFERENCES

- Asrul. (2015). *Evaluasi Pembelajaran*. Bandung, Citapustaka.
- Allanta, T. R., & Puspita, L. (2021). Analisis keterampilan berpikir kritis dan self efficacy peserta didik: Dampak PjBL-STEM pada materi ekosistem. *Jurnal Inovasi Pendidikan IPA*, 7(2), 158–170. <https://doi.org/10.21831/jipi.v7i2.42441>
- Aripin, I. (2017). Pengembangan soal-soal pilihan ganda untuk mengukur kemampuan berpikir kritis siswa pada konsep sistem regulasi manusia untuk jenjang SMA. *Mangifera Edu*, 2(1), 43–49. <https://doi.org/2622-3384>
- Bailin, S. (2002). Critical thinking and science education. *Science and Education*, 11(4), 361–375. <https://doi.org/10.1023/A:1016042608621>

- Chukwuyenum, A. N. (2013). Impact of critical thinking on performance in mathematics among senior secondary school students in Lagos State. *IOSR Journal of Research & Method in Education (IOSRJRME)*, 3(5), 18–25. <https://doi.org/10.9790/7388-0351825>
- Dharmawati, Rahayu, S., & Mahanal, S. (2016). Pengembangan instrumen asesmen berpikir kritis untuk siswa SMP kelas VII pada materi interaksi makhluk hidup dengan lingkungan. *Jurnal Pendidikan Pengembangan*, 1(64), 1598–1606.
- Dibattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, 2(2). <https://doi.org/http://dx.doi.org/10.5206/cjsotl-rcacea.2011.2.4>
- Duchscher, J. E. B. (1999). Catching the wave : understanding the concept of critical thinking. *Journal of Advanced Nursing*, 29(3), 577–583.
- Eriza, A. S., Selaras, G. H., Yogica, R., & Armen, A. (2020). Analisis pemahaman peserta didik tentang konsep materi keanekaragaman hayati menggunakan two-tier multiple choice test di SMAN 1 Rupert Utara. *Atrium Pendidikan Biologi*, 5(April), 48–55. <https://doi.org/2656-1700>
- Facione, P. A. (2015). *Critical thinking : What it is and why it counts* (7th ed.). Measured Reason LLC.
- Griffard, P. B., & Wandersee, J. H. (2001). The two-tier instrument on photosynthesis: What does it diagnose? *International Journal of Science Education*, 23(10), 1039–1052. <https://doi.org/10.1080/09500690110038549>
- Grigg, L. M., & Pomahac, G. U. Y. A. (2016). Faculty of Education , University of Calgary Critical Thinking in Science Education : Can Bioethical Issues and Questioning Strategies Increase Scientific Understandings ? Author (s): THELMA M . GUNN , LANCE M . GRIGG and GUY A . POMAHAC Source : The Jo. *The Journal of Educational Thought*, 42(May).
- Haka, N. B., Hamid, A., Dwi, A., Rudhini, M., & Riski, R. A. (2019). Pengembangan instrumen evaluasi two-tier multiple choice terhadap literasi sains berbantuan personal computer. *Biosfer: Jurnal Tadris Biologi*, 10(2), 201–214. <https://doi.org/2580-4960>
- Hala, Y., Arifin, A. N., Satar, S., & Saenab, S. (2019). Identification of biology student's misconception in Makassar State University on cell biology by applying two-tier MCQs Method. *Journal of Physics: Conference Series*, 1387. <https://doi.org/10.1088/1742-6596/1387/1/012004>
- Harahap, L. J., Ristanto, R. H., & Komala, R. (2020). Assessing critical thinking skills and mastery concept: The case of ecosystem material. *EDUSAINS*, 12(2), 223–232. <https://doi.org/http://doi.org/10.15408/es.v12i2.16544>
- Haslam, F., & Treagust, D. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education*, 21(3), 203–211. <https://doi.org/10.1080/00219266.1987.9654897>

- Husna, N. A., Zaini, M., & Winarti, A. (2021). The validity of biology module for senior high school on grade X in even semester based on critical thinking skills. *BIO-INOVED: Jurnal Biologi-Inovasi Pendidikan*, 3(1), 28–38. <https://doi.org/10.20527/bino.v3i1.9918>
- Ikhsanudin, & Subali, B. (2018). Content validity analysis of first semester formative test on biology subject for senior high school. *Journal of Physics: Conference Series*, 1097(1). <https://doi.org/10.1088/1742-6596/1097/1/012039>
- Kurniasih, N., & Haka, N. B. (2017). Penggunaan tes diagnostik two-tier multiple choice untuk menganalisis miskonsepsi siswa kelas X pada materi archaeobacteria dan eubacteria. *Biosfer: Jurnal Pendidikan Biologi*, 8(1), 114–127.
- Laksono, P. J. (2018). Pengembangan dan penggunaan instrumen two-tier multiple choice pada materi termokimia untuk mengukur kemampuan berpikir kritis. *Orbital: Jurnal Pendidikan Kimia*, 2(2), 80–92.
- Magdalena, I., Fauzi, H. N., & Putri, R. (2020). Pentingnya evaluasi dalam pembelajaran dan akibat memanipulasinya. *Bintang: Jurnal Pendidikan Dan Sains*, 2(2), 244–257. <https://ejournal.stitpn.ac.id/index.php/bintang>
- Ngafifah, S. (2020). Penggunaan google form dalam meningkatkan efektivitas evaluasi pembelajaran daring siswa pada masa COVID19 Di SD IT Baitul Muslim Way Jepara. *As-Salam: Jurnal Studi Hukum Islam & Pendidikan*, 9(2), 123–144. <https://doi.org/10.51226/assalam.v9i2.186>
- Noviana, M., Sajidan, S., & Pugh, P. (2014). Pengembangan instrumen evaluasi two-tier multiple choice question untuk mengukur keterampilan berpikir tingkat tinggi pada materi kingdom plantae. *Jurnal Inkuiri*, 3(2), 60–74. <https://doi.org/2252-7893>
- Rizal. (2017). Mengajar cara berpikir, meraih ketrampilan abad 21. *Seminar Nasional Pendidikan PGSD UMS & HDPGSDI Wilayah Jawa Pendidikan*, 390–406. <http://hdl.handle.net/11617/9134>
- Rukayah, R., Atmojo, I. R. W., Hartono, H., & Daryanto, J. (2020). Conceptual biotechnology measured by using a two-tier multiple test. *Advances in Social Science, Education and Humanities Research*, 397(Icliqe 2019), 565–569.
- Sari, K., & Paidi. (2019). Metacognitive knowledge and critical thinking biology 11 th of public senior high school in Bogor. *IOP Publishing*, 1241, 1–6. <https://doi.org/10.1088/1742-6596/1241/1/012056>
- Sesli, E., & Kara, Y. (2012). Development and application of a two-tier multiple-choice diagnostic test for high school students understanding of cell division and reproduction. *Journal of Biological Education*, 46(4), 37–41. <https://doi.org/10.1080/00219266.2012.688849>
- Simpson, E., & Courtney, M. (2002). Critical thinking in nursing education: Literature review. *International Journal of Nursing Practice*, 8, 89–98.
- Sudirman, Halima, & Hidayat, M. Y. (2021). Implementation of guided inquiry learning model assisted by three tier test on critical thinking. *Jurnal Pendidikan Fisika*, 9(2). <https://doi.org/2550-0325>
- Suryawirawati, I. G., Ramdhan, B., & Juhanda, A. (2018). Analisis penurunan

- miskonsepsi siswa pada konsep pemanasan global dengan tes diagnostik (Two-tier test) setelah pembelajaran Predict-Observe--Explain (POE). *Journal of Biology Education*, 1(1), 93–105.
- Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, December 1986, 119–207. <https://doi.org/10.1007/BF02356835>
- Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, 32(8), 37–41. <https://doi.org/10.1080/09500690902951429>
- Walid, A., Marfhadella, P., & Satria, I. (2021). Development of an assessment to measure science process skills for the interaction of living things with their environment in junior high school. *Science Education and Application Journal*, 3(2), 113. <https://doi.org/10.30736/seaj.v3i2.427>
- Yamtinah, S., Saputro, S., & Utami, B. (2016). Content validity and scoring of two tier as measuring instrument of science process skills for knowledge aspects in chemistry learning. *Prosiding ICTTE FKIP UNS*, 1(1), 911–916. <https://doi.org/2502-4124>
- Zahara, N. (2015). Evaluasi pembelajaran online berbasis web sebagai alat ukur hasil belajar siswa pada materi dunia tumbuhan kelas X MAN Model Banda Aceh. *“Evaluasi Pembelajaran Online Berbasis Web Sebagai Alat Ukur Hasil Belajar Siswa Pada Materi Dunia Tumbuhan Kelas X Man Model Banda Aceh*, 53(9), 1689–1699.
- Zein, A., Fadillah, M., & Novianti, R. (2013). Hubungan antara validitas butir , reliabilitas , tingkat Kesukaran dan daya pembeda soal ujian semester genap bidang studi biologi kelas XI SMA / MA Negeri di Kota Padang tahun pelajaran 2010 / 2011 *. *Prosiding Semirata FMIPA Universitas Lampung*, 2009.