



Construction of biology critical thinking test of high school students

Taufiq Satria Mukti^{1,*}, Melly Elvira², Fantika Febry Puspitasari³



Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia

¹ tsatriam@uin-malang.ac.id *; ² melly@uin-malang.ac.id; ³ fantika@mpi.uin-malang.ac.id

* Corresponding author

ARTICLE INFO

Article history

Received June 4, 2020
Revised November 5, 2020
Accepted November 16, 2020

Keyword:

Critical thinking
Polytomous scoring
Partial credit model
Reasoned multiple choice

ABSTRACT

Students must have various abilities, one of which is critical thinking skills. Critical thinking is an ability that is an indicator of successful learning and is directly related to real life. Should carry out the development of critical thinking skills effectively and efficiently in education. Critical thinking skills of students, accurate measurement tools are needed. The measuring instrument is in the form of a test instrument developed through research. Conduct the researched with a quantitative approach by describing the characteristics of the critical thinking test instrument. The research method was the development of a test with the research subjects of class X MIPA students from seven public high schools in Kendal Regency with the criteria of high, medium, and low-ranking schools. Developed the test in the form of multiple-choice reasoned with three scoring categories for each item according to the scoring of the polytomous items in the PCM 1-PL model (Partial Credit Model 1 Parameter Logistic). The feasibility test of the critical thinking test instrument was carried out with experts in measurement, assessment, and Biology Learning and was tested empirically on 1118 students. The research data were analyzed using the EXCEL, SPSS 16, QUEST and PARSCALE applications. The results showed that: (1) The test instrument with four categories of scores fit the PCM 1-PL model; (2) the test instrument has a high-reliability value; (3) The test instrument has a right level of difficulty; (4) The test instrument can be used to measure the ability of students in the range of -3.7 to 2.9. The test instrument developed was feasible to be used to measure students' critical thinking skills with three categories of polytomous scoring with the PCM 1-PL model.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Introduction

Critical thinking is a combination of mental processes, strategies and a person's representations that use to solve problems, make decisions, and learn new concepts (Sternberg, 1986). This opinion is also

confirmed by Wilson (2004) that critical thinking is also a unity of understanding concepts and complex ways of thinking. The agreement of the idea of critical thinking is a combined representation. It is related to knowledge which can be in the form of concepts, opinions, or ideas in

humans to find reasonable and reflective thinking patterns to responses in the way of decisions or conclusions to take various actions on what is believed (Arends, 2012; Ennis, 2011; Moon, 2007).

In the context of learning, especially biology, Moon (2007) states that critical thinking is the result of training and habituation of daily, continuous learning activities so that complex thinking processes will occur. In the learning process and real life, critical thinking can be in the form of the ability to respond to and solve problems related to real biological objects (Addy et al., 2014). In line with that, following the nature of biology as part of science the learning process carried out by the scientific method must involve various skills, especially critical thinking to solve problems (Towle & Twole, 1989). Therefore, we need action to detect critical thinking skills in students.

Detecting critical thinking skills according to York et al. (2015) needs to be done through a process of measuring students because this ability is an essential ability that can use as an indicator of the success of the learning process that has reached competency standards. Besides, according to the opinion (Paul & Elder, 2005), the need for critical thinking in learning is to get used to complex thinking to conceptualize, synthesize or evaluate information obtained through observation, experience, reasoning or reflection. Measure critical thinking in learning through a test. Critical thinking tests can also be used as a form of training and habituation of ways of thinking to face and solve problems in real life (Chiras, 1992; Palm, 2008) so that based on this description. Critical thinking tests need to develop because in essence Biology learning is very close to real problems that exist in everyday life that require various responses to find solutions to all problem conditions (Hidayati, 2016).

Measuring critical thinking skills is generally done through a measurement process using a test instrument in the form of a description. However, essay questions have weaknesses as described in the study (McPeck, 2016), test in the way of opinions had faults and problems that cannot tolerate in the assessment process. This problem is the existence of an element of subjectivity from the assessors when examining the test results or scoring the responses and affect the measurement

results that do not reflect the actual products. Also, the process of scoring takes a long time is also a weakness of this test. The implementation of the assessment of critical thinking skills in the education sector is still in the low category, which is around 45% (Lane, 2015). The results of research conducted Huber and Kuncel (2016) show that the implementation of the assessment of critical thinking skills takes a relatively long time when compared to the evaluation of learning outcomes in general. The length of time is the whole from planning to implementation. Based on the results of other research, the assessment of critical thinking ability still focused on the results of achieving competency standards in certain subjects (Anisa, 2017). Thus, not many researchers have reviewed the characteristics of the instruments used to estimate critical thinking skills.

Lack of researchers who measure and review critical thinking skills with a variety of test models, it is necessary to use a new test model which is assumed to use in estimating abilities accurately. So far, critical thinking skills are rarely measured using a multiple-choice test model. It is influenced by the guessing factor in implementation (Osterlind, 1998) and requires particular expertise in compiling the test items to be used. However, research Hartini and Sukardjo (2015), Akbar et al. (2017) found that multiple-choice tests can measure critical thinking skills. In physics, similar research was conducted by Putri and Istiyono (2017) who found that the critical thinking skills of high school students in Physics subjects can be measured using multiple-choice tests, namely by modifying the regular multiple-choice model. The model is a reasoned multiple choice. In the field of Biology, can also estimate the ability to think critically with a similar test model (Mukti & Istiyono, 2018).

The reasoned multiple-choice test model that developed E Istiyono et al. (2014) with a polytomous scoring model with four score categories. In this model, based on modern measurement theory (item response theory), analysis with PCM (Partial Credit Model) can be carried out (Hambleton et al., 1991). In the PCM model, the logistical parameter 1 PL means that the calibration process only includes the difficulty level of the questions based on the response, and does not consider the

answer roughly as suggested by [Adams and Khoo \(1996\)](#). This model is also known as the tau model or other forms of development in the RM model (Rasch Model) ([Hambleton et al., 1991](#)). The results of the study [Edi Istiyono et al. \(2014\)](#) state that the form of the test is sufficient for measuring and describing higher-order thinking skills and critical thinking skills. That is because the thinking process involves knowledge and the ability to solve complex problems. Critical thinking aspects developed in the test indicators include assumptions, arguments, analysis, evaluation, and conclusions.

Based on the description above, the researchers developed a measuring tool that can estimate the critical thinking ability of SMA Negeri students in Biology Class X Mathematics and Natural Sciences in Kendal Regency. Data analysis uses IRT (Item Response Theory) analysis to describe the characteristics of the instrument which include: model fit (Goodness of Fit Test), test reliability, difficulty level, and test instrument information function based on existing criteria. The PCM 1-PL model used involved three scoring categories. That is based on differences in the concept of scoring in predecessor researchers conducted by [E Istiyono et al. \(2014\)](#). In the field of physics, can mistake the completion of items using the counting method for an answer with a pattern of wrong answer responses and right reasons. That happens because of miscalculations or inaccurate factors. However, in Biology which is emphasized by [B Subali \(2009\)](#) that misconceptions are not possible when students can provide reasons, but the answers are wrong, this confirms that there is no guessing element in the scoring model used.

Method

The test development model adopted from [Mardapi \(2016\)](#) consists of compiling test specifications, writing tests, content validity, testing, examining instrument characteristics based on empirical data, and taking measurements. This article analysis of the characteristic of the instrument on the response data of State Senior High School students in Biology Class X MIPA in Kendal Regency. The quality of the instrument is based on the information from the analysis of the characteristics of the instrument.

Developing quality instruments requires detailed and specific developmental stages to obtain information capable of describing the actual conditions.

The first stage of test development is compiling test specifications based on determining the competencies to be tested. The test is a competency in the subject of Biology subject in class X MIPA SMA Negeri 1 semester which consists of 1) Biological Scope, 2) Scientific Methods, 3) Classification, 4) Biodiversity, 5) Fungi, 6) Protists, 7) Bacteria, 8) Viruses. Meanwhile, the critical thinking theory used includes: Aspect 1 (A1) Assumptions, sub-aspect 1 (SA11): determining relevant hypotheses, sub-aspect 2 (SA12): determining the results of considerations based on background and facts; Aspect 2 (A2): Argument, sub-aspect 1 (SA21): make arguments based on facts and knowledge, sub-aspect 2 (ASA22): Identify a cause and effect/ reason relationship; Aspect 3 (A3): Analysis, sub-aspect 1 (SA31): Analysis of the background and purpose of information, sub-aspect 2 (SA32): linking information to human activities; Aspect 4 (A4): Evaluation, sub-aspect 1 (SA41): checking the suitability of a problem with a solution, sub-aspect 2 (SA42): making criticism of a problem; Aspect 5 (A5): Summing up, Sub-aspect 1 (SA51): Inducing informed thinking, sub-aspect 2 (SA52): Deducing thinking based on information. Based on the determination of competence, it was then compiling test specifications suitable for use.

Test specification consists of the form of the test and the length of the test/length of the test. The format of the test is a reasoned multiple choice. That is observing students' critical thinking skills through the answers and reasons they choose. The development of critical thinking tests consists of 45 items with A and B tests, each of which is 25 points (5 anchors) with a duration of 90 minutes or the equivalent of 2 hours of Biology subjects. For each item, there are five answer choices and five choices of reasons so that the scoring of the items is 1-3 scoring.

The second stage is determining material in Biology competence and writing test items. Items distribution refers to the highest percentage of items based on essential material and material on the competency standards of Biology class X material graduates. The matrix serves as control over the suitability of test items to

aspects, sub-aspects and material of Biology. The selected material then becomes a reference for making the grid.

The third and fourth stages are the stages of arranging the grid and arranging the items that refer to the grid. The items consist of two sets A and B with each group consisting of 20 items and five anchor items. The available answer choices are multiple-choice with reasonable answer choices. The choice of reasoned answers aims to be able to analyze the critical thinking skills of students.

The fifth stage of test development is to determine the form of scoring. The multiple-choice model argues that it uses a polytomous scoring model which functions to determine the critical thinking ability in each category of answer choices. Analysis of the response data of students' answers using the PCM model (Partial Credit Model) 1-PL. The scoring provisions in each category are: Category-1 if the answer is wrong and the reason is wrong with a score of 0; Category-2 if the answer is correct and the reason is wrong with a score of 1; Category-3 if the answer is correct and the reason is correct with a score of 2.

The sixth stage is the validation stage. Content validity by expert judgment consisting of measurement, assessment, Biology Learning experts, and teachers. Content validity by looking at the suitability of competencies, indicators and items from the Biology material in the critical thinking test instrument. The purpose of content validity is to determine the feasibility of the items in terms of concept, construction, language and the effectiveness of the items to measure the level of students' critical thinking skills. The sixth stage is the validation stage. Content validity by expert judgment consisting of measurement, assessment, Biology Learning experts, and teachers. Content validity by looking at the suitability of competencies, indicators and items from the Biology material in the critical thinking test instrument. The purpose of content validity is to determine the feasibility of the items in terms of concept, construction, language and the effectiveness of the items to measure the level of students' critical thinking skills.

The implementation of a wide-scale trial of the test subjects was as many as 1118 students of class X MIPA Public Senior High School in Kendal Regency who met the criteria for the ability of students

from schools with low, medium and high categories. The selection is based on the 2017 National Exam Score ranking in the Biology subject for Public High Schools. The implementation of the test uses test sets A and B with the design of seats for front, back, right, and left students working on the questions differently and alternating with the code questions A and B, as illustrated in the following [Figure 1](#).

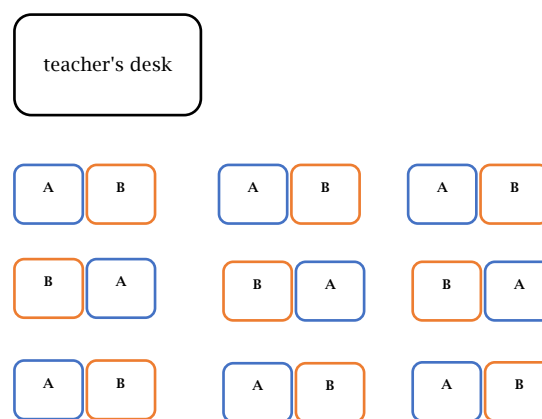


Figure 1. Seating design in test

The last stage is to collect data and analyze data. Data analysis of test development results in the form of student responses with the help of Excel, SPSS, Quest, and Parscale application programs is in the form of model fit (goodness of fit test), difficulty level, reliability, and test information functions.

Results and Discussion

Results of the estimation of the suitability of the test instrument model (goodness of fit test) are presented in [Table 1](#).

Table 1. Estimated Test Item Parameters

Criterion	Item Estimation
Average and standard deviation	-0.14 ± 0.26
Reliability	0.93
MNSQ INFIT mean and standard deviation	1.01 ± 0.04

Estimation of test parameters with the QUEST application program is in a suitable category if it meets the criteria according to [Adams and Khoo \(1996\)](#) by looking at the average MNSQ Infit value (Infit Mean of Square) and its standard deviation value. The fit score is in the MNSQ Infit range ± 1.00 and the default deviation is 0.00. The acquisition of the MNSQ Infit value is 1.0,

and the standard deviation value is 0.04 so that all items on the test fit for the PCM 1-PL model. Criteria for the fit of the test instrument items (goodness of fit test) on each test item can be seen in Figure 2 which states that all items are in the value range of 0.73 to 1.30 so that all items are fit and fit for use.

Another characteristic of the instrument's feasibility is the reliability index. The reliability criteria on the instrument serve as an indicator of the stability of the test, describing the reliability value of the test so that the measurement results can provide information, make decisions, accurate (Hambleton et al., 1991). The instrument

reliability estimates of 0.93 (Guilford, 1956) states that the reliability index obtained is in the high category and the test instrument is very suitable to be used to measure students' critical thinking skills.

The PCM 1-PL model of the test instrument needs to pay attention to the difficulty level of the test items. This test item is an assumption in the implementation of tests that can measure the ability of students by covering all levels of students' ability both high, medium and low (Hambleton et al., 1991). The difficulty level of the test instrument base on the test results on the following histogram (Figure 2).

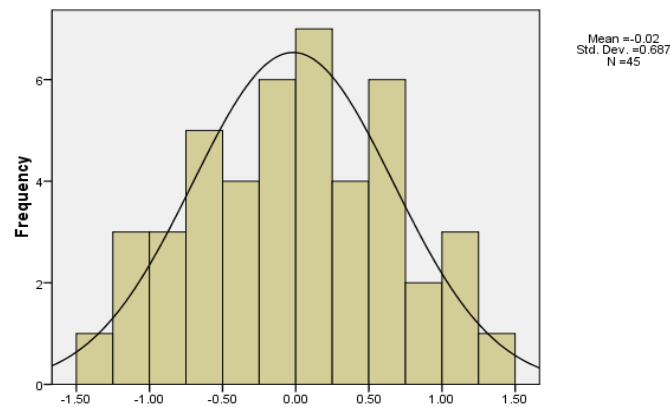


Figure 2. Distribution of difficulty levels

The distribution of the difficulty level of the test instrument is close to the normal distribution. However, there is no requirement that one test must have a difficulty level with the distribution following the normal distribution (Bambang. Subali, 2011). Thus, the test results of the test instrument meet the requirements and are declared feasible according to Hambleton et al. (1991) because it has a difficulty level range of -2.00 to 2.00.

Table 2. Level of difficulty in aspects and sub-aspects

A	Average	SA	Average
A1	-0.15	SA11	0.12
		SA12	-0.42
A2	0.10	SA21	-0.04
		SA22	0.24
A3	-0.05	SA31	-0.16
		SA32	0.06
A4	0.04	SA41	0.05
		SA42	0.03
A5	0.14	SA51	0.29
		SA52	-0.01

Information: A (aspect), SA (Sub-aspect)

The matrix in Table 2 serves to describe the level of difficulty of the critical thinking test instrument developed by the following aspects, sub-aspects of critical thinking, and biology material.

Table 2 shows the average level of difficulty of the distribution of test items according to aspects, sub-aspects and material of Biology. Still, the emphasis on the level of difficulty is the distribution of items that state the level of difficulty in critical thinking aspects and sub-aspects. The highest level of difficulty in the critical thinking aspect is at A5 (Concluding), and the lowest level of difficulty is at A1 (Assumption). Meanwhile, the sub-aspect with the highest difficulty level was SA51 (Inducing thought based on information), and the lowest was SA12 (determining considerations based on background and facts).

Analysis with the PCM 1-PL model is to follow a partial credit, and if the individual's higher abilities will have a score in the high stage category as well

(Widhiarso, 2010). The percentage of item responses with score categories 1, 2, and 3 on the test results of the test instrument is in Table 3.

Table 3. Percentage of answer in the score category

A	SA	Score category (%)		
		1	2	3
A1	SA11	41.49	14.65	44.00
	SA12	14.85	7.11	78.04
A2	SA21	25.89	13.27	58.87
	SA22	54.20	11.03	34.78
A3	SA31	28.08	5.06	66.86
	SA32	37.89	15.95	46.15
A4	SA41	43.08	11.45	45.47
	SA42	42.19	6.74	51.07
A5	SA51	53.14	9.55	37.31
	SA52	35.79	10.99	53.21

Information: A (aspect), SA (Sub-aspect)

The interpretation of Table 3 is that the Partial Credit pattern in the test instrument has characteristics that include a moderate level of difficulty, this evidenced by the total percentage between categories 1, 2 and 3 does not occur significantly.

The test information function on the results of the analysis serves as a form of strengthening the test parameter estimates for the reliability value (Retnawati, 2015). In the analysis of IRT, the information function and SEM (Standard Error of Measurement) also apply to explain the functionality, stability and accuracy of the tests used by calibrating the item responses to the abilities of students on a logit scale. The results of the test information function estimation based on the test instrument trials in Figure 3.

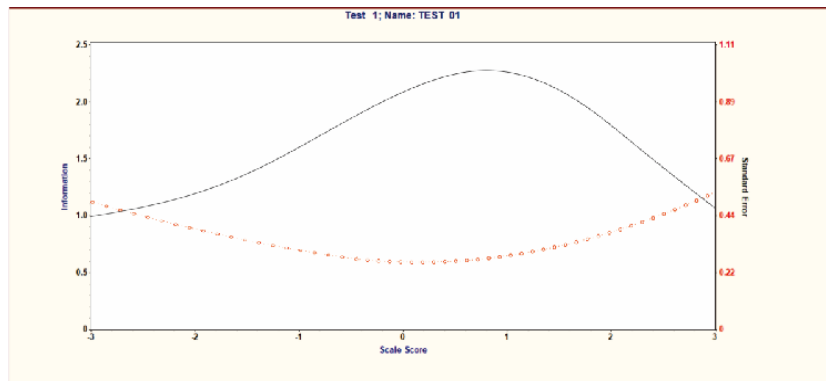


Figure 3. Information functional and measurement error standards

Response analysis with Parscale shows that in Figure 3, the critical thinking test instrument can measure the ability of students with a range of ability levels of -2.7 to 2.9. These results conclude that the

test instrument is excellent and feasible because it can measure students' high or lowest ability according to the response data of the test results.

31. Flagellates are parasitic protozoa on red, white blood cells and liver cells. The presence of flagellates in freshwater is an indicator of water pollution. The life phase is flagella in the extracellular phase and not in the intracellular phase. The impact is elephantiasis caused by Trypanosoma through a mosquito bite intermediary host.

The information to be conveyed is ...

- The existence of Flagellates has a beneficial role for humans.
- It needs deep environmental concern to create a healthy environment.
- Protozoa have a role as indicators for uncontaminated waters.
- Trypanosoma is a group of protozoa that has a negative role on humans.
- Must be wise in using water to make ends meet.

Reason:

- Concern for the environment must start from itself to make the environment healthy.
- Trypanosoma plays a positive role because it functions as a drug.
- Trypanosoma plays a negative role because it can cause health problems in humans.
- Trypanosoma is very unlikely to live in the environment around us.

Figure 4. Item example (translate version)

Figure 4 example of test items to measure critical thinking skills in Biology Class X Semester 1. One indicator that students have the ability to think critically is being able to deduce information from the substance/content of learning according to the competencies to be achieved. This problem sharpens the ability to think critically about biological material (Protozoa) as well as examining information about its application in life. The answer to the example problem is D for the answer and C for the reason.

Conclusion

The test instrument fulfils the requirements to measure the critical thinking skills of students of SMA Negeri Class X MIPA in Semester 1 Biology subject. Two sets of test instruments (Test A and Test B) with the number of items 25 with five anchor points fulfil empirical evidence of model suitability (goodness of fit test) on the PCM model (Partial Credit Model) based on the polytomous score of four categories. The test instrument has a right level of difficulty with a value range of -2.00 and 2.00. The test instrument has a high-reliability value of 0.86, so it qualifies as a useful measuring tool. The test instrument can measure critical thinking skills with the respondent's ability range of -3.7 to 2.90.

References

- Adams, R. J., & Khoo, S. T. (1996). *Quest: The interactive test analysis system version 2.1*. The Australian Council for Educational Research.
- Addy, T. M., LePrevost, C., & Stevenson, M. (2014). Thinking critically in undergraduate biology: Flipping the classroom and problem-based learning. *Double Helix: A Journal of Critical Thinking and Writing*, 2(1), 1-9. <https://doi.org/10.37514/DBH-J.2014.2.1.09>
- Akbar, M. N., Firman, H., & Rusyati, L. (2017). Developing science virtual test to measure students' critical thinking on living things and environmental sustainability theme. *Journal of Physics: Conference Series*, 812, 012106. <https://doi.org/10.1088/1742-6596/812/1/012106>
- Anisa, A. (2017). Meningkatkan keterampilan berpikir kritis peserta didik melalui pembelajaran IPA berbasis potensi lokal Jepara. *Jurnal Inovasi Pendidikan IPA*, 3(1), 1-11. <https://doi.org/10.21831/jipi.v3i1.8607>
- Arends, R. (2012). *Learning to teach* (9th ed.). McGraw-Hill.
- Chiras, D. D. (1992). Teaching critical thinking skills in the biology & environmental science classrooms. *The American Biology Teacher*, 54(8), 464-468. <https://doi.org/10.2307/4449551>
- Ennis, R. (2011). Critical thinking. *Inquiry: Critical Thinking Across the Disciplines*, 26(2), 5-19. <https://doi.org/10.5840/inquiryctnews201126215>
- Guilford, J. P. (1956). *Fundamental statistic in psychology and education* (3rd ed.). McGraw-Hill.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hartini, H., & Sukardjo, S. (2015). Pengembangan higher order thinking multiple choice test untuk mengukur keterampilan berpikir kritis IPA kelas VII SMP/MTs. *Jurnal Inovasi Pendidikan IPA*, 1(1), 86-101. <https://doi.org/10.21831/jipi.v1i1.4535>
- Hidayati, N. (2016). Hasil belajar dan keterampilan berpikir kritis siswa Madrasah Tsanawiyah dalam pembelajaran IPA melalui kerja ilmiah. *Proceeding Biology Education Conference*, 118-127. <https://jurnal.uns.ac.id/prosbi/article/view/5672>
- Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, 86(2), 431-468. <https://doi.org/10.3102/0034654315605917>
- Istiyono, Edi, Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PysTHOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 18(1), 1-12. <https://doi.org/10.21831/pep.v18i1.2120>

- Istiyono, E, Mardapi, D., & Suparno, S. (2014). Effectiveness of reasoned objective choice test to measure higher order thinking skills in physics implementing of curriculum 2013. *International Conference on Educational Research and Evaluation*, 101-110. <https://repository.bbg.ac.id/handle/399>
- Lane, S. (2015). Performance assessment. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the Bubble Test* (pp. 131-184). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781119210863.ch5>
- Mardapi, D. (2016). *Pengukuran, penilaian dan evaluasi pendidikan*. Nuha Litera.
- McPeck, J. E. (2016). *Critical thinking and education*. Routledge.
- Moon, J. (2007). *Critical thinking: An exploration of theory and practice* (1st ed.). Routledge. <https://doi.org/10.4324/9780203944882>
- Mukti, T. S., & Istiyono, E. (2018). Multiple-choice with reason (MCR) to measure the critical thinking skill on natural science: plant and animal movement system. *5th ICRiems Proceedings*, 35-40. [http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/SE06_Taufiq Satria Mukti.pdf](http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/SE06_Taufiq%20Satria%20Mukti.pdf)
- Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other format* (Seconds). Kluwer Academic Publishers.
- Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research, and Evaluation*, 13, Article 4. <https://doi.org/10.7275/Qqpc-ws45>
- Paul, R., & Elder, L. (2005). *A guide for educators to critical thinking competency standards: Standards, principles, performance indicators, and outcomes with a critical thinking master rubric*. [https://www.criticalthinking.org/resources/PDF/CT-competencies 2005.pdf](https://www.criticalthinking.org/resources/PDF/CT-competencies%202005.pdf)
- Putri, F. S., & Istiyono, E. (2017). The development of performance assessment of stem-based critical thinking skill in the high school physics lessons. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL & SCIENCE EDUCATION*, 12(5), 1269-1281. <http://www.ijese.net/makale/1894.html>
- Retnawati, H. (2015). The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing. *Turkish Online Journal of Educational Technology*, 14(4), 135-142. <http://www.tojet.net/volumes/v14i4.pdf#page=144>
- Sternberg, R. J. (1986). *Critical thinking: Its nature, measurement, and improvement*. National Inst. of Education. <https://eric.ed.gov/?id=ED272882>
- Subali, B. (2009). Pengembangan tes pengukuran keterampilan proses sains pola divergen mata pelajaran biologi SMA. *Prosiding Seminar Nasional Biologi, Lingkungan Dan Pembelajarannya*, 581-593. [http://staff.uny.ac.id/sites/default/files/penelitian/Bambang Subali, Dr/semnas bio _Bambang Subali_UNY-2009.pdf](http://staff.uny.ac.id/sites/default/files/penelitian/Bambang%20Subali,%20Dr/semnas%20bio%20_Bambang%20Subali_UNY-2009.pdf)
- Subali, Bambang. (2011). Pengukuran kreativitas keterampilan proses sains dalam konteks assessment for learning. *Jurnal Cakrawala Pendidikan*, 30(1), 130-144. <https://doi.org/10.21831/cp.v1i1.4196>
- Towle, A., & Twole, A. (1989). *Modern biology*. Holt Rinehart & Winston.
- Widhiarso, W. (2010). *Estimasi abilitas individu dengan metode maximum likelihood dalam pemodelan IRT*. Repository Universitas Gajah Mada. <https://repository.ugm.ac.id/275984/>
- Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge. <https://doi.org/10.4324/9781410611697>
- York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research, and Evaluation*, 20, Article 5. <https://doi.org/10.7275/hz5x-tx03>