



Content Validity of the Critical Thinking Skill Test Instrument on Computer Based Test (CBT) Ecology Lesson for High School Education

Irma Kharisma¹, Andi Ulfa Tenri Pada^{2*}, Supriatno², Safrida², Ismul Huda²

¹ Magister of Biology Education, Faculty of Teacher Training and Education, Universitas Syiah Kuala, Banda Aceh, Indonesia.

² Department of Biology Education, Faculty of Teacher Training and Education, Universitas Syiah Kuala, Banda Aceh, Indonesia.

Received: January 10, 2024

Revised: April 17, 2024

Accepted: June 25, 2024

Published: June 30, 2024

Corresponding Author:

Andi Ulfa Tenri Pada

andi_ulfa@usk.ac.id

DOI: [10.29303/jppipa.v10i6.6914](https://doi.org/10.29303/jppipa.v10i6.6914)

© 2024 The Authors. This open access article is distributed under a (CC-BY License)



Abstract: Research to measure students' Critical Thinking Skills has been widely carried out in Indonesia. However, the development of Critical Thinking Skill measuring tools in Ecology using CBT is still limited. This study is part of research and development to develop a good instrument product for measuring students' Critical Thinking Skills. The question preparation format uses a two tier multiple-choice form. This study aims to prove the content validity of the Critical Thinking Skill instrument in Ecology Subject for Senior High School. Scale using Likert model and multiple-choice model with content validity coefficient based on expert assessments with Aiken's formula. There are three experts who assess the items' relevancy, construction, and clarity using indicators of both scale formats. The results of the expert assessments are then used to calculate the coefficient of the validity with Aiken formula. The results showed that the content validity coefficient based on expert assessment on Likert format with Aiken formula is at 0.75-1.00 for each, while using the Aiken formula.

Keywords: Aiken formula; Critical thinking skill test; Validity coefficient

Introduction

Currently, 21st century skills integrated with information and communication technology (ICT) have become a global competency goal (Istiyono et al., 2020). Furthermore, Hidayah et al. (2020) stated that in the 21st century, all professions think that Critical Thinking Skills are very important to develop. Therefore, educational practitioners are now trying to develop students' Critical Thinking Skills to prepare their graduates to face 21st century competition. Based on the facts above, what exactly are Critical Thinking Skills? Ennis (n.d.) defines Critical Thinking Skill as a thinking skill to compare two or more pieces of information. Critical thinkers are considered to be able to conclude a problem with full consideration and can re-evaluate the decisions that have been taken. Apart from that, critical thinking is used in the basic process of thinking to

analyze opinions and provide ideas based on logical reasoning.

In line with Ennis, Kennedy et al. (2013) mentions several Critical Thinking Skills, namely: recognizing problems; finding ways that can be used to solve problems; collect and compile the necessary information; understand and use appropriate language, analyze data, assess facts, and evaluate statements; recognize logical relationships between problems; draw necessary conclusions and equations; examine similarities and conclusions.

Critical thinking skills cannot be directly possessed by students but are obtained through practice. Science learning requires educators to generate questions that challenge students' ideas, knowledge and assumptions to connect information and build understanding to help students engage in higher level thinking and learning (Gillies et al., 2012). Assessment tools as a tool for measuring student learning success need to support

How to Cite:

Kharisma, I., Pada, A. U. T., Supriatno, Safrida, & Huda, I. (2024). Content Validity of the Critical Thinking Skill Test Instrument on Computer Based Test (CBT) Ecology Lesson for High School Education: English. *Jurnal Penelitian Pendidikan IPA*, 10(6), 3533-3540. <https://doi.org/10.29303/jppipa.v10i6.6914>

aspects of measuring critical thinking skills. The Critical Thinking Skill-based assessment process in implementing science learning has not yet emerged, while the critical thinking learning process in schools in Aceh is still very low (Ritonga et al., 2020). The test instrument does not contain questions that support critical thinking skills.

Critical thinking skills are the main and first thing that must be considered in learning in the era of industrial revolution 4.0, as in the learning framework developed by the Partnership for 21st Century Learning (Saleh, 2019). Although critical thinking is often referred to as the most important element in science learning, in reality critical thinking assessment is an area that is often neglected (Mundilarto, 2002). In fact, this assessment is very necessary to find out whether students have sufficient biological critical thinking skills and what treatment should be given (Elliot et al., 2019). One of the reasons why critical thinking assessment in science learning is often neglected is the lack of available tests that specifically measure critical thinking skills in biology. Problems like this have actually been frequently complained about by several researchers (Istiyono et al., 2014). This means that the development of standardized critical thinking tests must be adapted to existing needs (biology) so that similar problems do not always recur.

Another problem that arises is the test format which is less appropriate in measuring critical thinking skills and accommodating material coverage. The forms of tests that are often used are essays and multiple choice. Essays are the most comprehensive form of tests (Khan, 2017). However, it has many weaknesses, such as limited measurement of high-level thinking, more time-consuming and expensive, high subjectivity, and difficult to determine its validity and reliability (Shaaban, 2014). Meanwhile, multiple choice has many advantages, such as being easy to apply in large classes, a high level of objectivity, broad material coverage, and can be corrected easily (Johnson et al., 2014). The weakness of multiple choice is that it is less comprehensive and students' thinking processes cannot be seen clearly (Istiyono, 2014). Based on these advantages and disadvantages, an appropriate form of test must be determined by considering validity, reliability and completeness. The solution to this problem is by implementing a two-level multiple choice test (Winarti et al., 2017). This format consists of several answer choices and several reason choices, so it will require students to think in determining reasons that match their answer choices (Istiyono et al., 2019).

This is in accordance with what was stated in Ennis (1985) that to measure students' Critical Thinking Skills, both at the individual and group level, effective assessment techniques are needed. There are several assessment techniques that can be used to measure

Critical Thinking Skills. Ennis (2011) believes that modified tests in the form of reasoned multiple choice tests are believed to be effective in measuring students' Critical Thinking Skills.

The reasoned multiple choice test (two-tier multiple choice) is an alternative test as a measuring tool that comprehensively covers science learning (Winarti et al., 2017). Cullinane et al. (2011) show that there is a modified multiple choice test, namely two-tier multiple-choice by including reasons at the second level of the answer which can be used to measure critical thinking abilities. Reasoned multiple choice tests require reasons that match the answer choices (Putri et al., 2016). A reasoned multiple choice test will effectively and fairly be able to carry out a more accurate assessment of science learning. Ennis (1993) also recommends a reasoned multiple choice test (two-tier multiple-choices) as a way to measure students' critical thinking skills. This type of test is considered to provide more sources of information about students' abilities. Syaifuddin et al. (2022) revealed that the first level answer choices in the two-tier assess students' descriptive or factual knowledge about phenomena. At the second level, analyze students' reasons for choosing their options at the first level.

Paying attention to the root of the problem, it is necessary to think how to solve it. Moreover, the application of the curriculum in higher education focuses on 21st century skills that often called the 4C skills integrated to critical thinking, creative thinking, communication, and collaborative. We offer a solution by developing assessment tool that can measure student's critical thinking skill. This assessment is expected to improve the critical thinking skill of the student. The student reasoning ability will be directed to produce arguments based on their concept understanding in Ecology lesson.

After developing the assessment, we need to prove whether the assessment tool has been optimally constructed to evaluate the quality of the assessment. The most important consideration in evaluating the quality of the test as a measurements tools is validity. Messick defines validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Stanley et al., 2009). From the interviews with practitioners in the educational fields, some practitioners question the validity of the questionnaire with Likert model in multiple choice models. Each practitioner has its own reasonable arguments. The Likert questionnaire model is easy to make and easy to read by the respondents, but the data obtained contain desirability bias. The multiple-choice questionnaire model is difficult to make and the

respondents need time to read, but more valid data can be obtained from it. Related to this problem, this study describes the proof of the content validity from the questionnaire in Likert and multiple-choice model with stratified scoring.

There are various opinions on the validity of the instruments used for the measurement, both in education and psychology. According to American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) in the Standards for Educational and Psychological Testing, validity refers to the degree of facts and theories that support the interpretation of instrument scoring, and the most important consideration in the development of an instrument (1999). Other experts point out that the validity of a measuring instrument is to what extent the measuring instrument able to measure what should be measured (Allen et al., 2001; Nunnally, 1978; Smith, 2005).

Meanwhile, Linn et al. (1995) explain that validity refers to the adequacy and interpretation appropriateness made of assessment, related to a specific use. This opinion is reinforced by Messick (1989) who writes that validity is an integrated evaluative policy concerning what extent of empirical facts and theoretical reasons support the adequacy and appropriateness of inferences and actions based on test scores or scores of an instrument. Based on those opinions, it can be concluded that validity will show supports to empirical facts and theoretical reasons for the interpretation of test scores or score of an instrument, and it is associated with the measurement precision.

There are three types of validity, namely: (1) criterion validity (criterion-related validity), (2) content validity, and (3) construct validity (Allen et al., 2001; Nunnally, 1978; Pozo-Muñoz et al., 2000; Sandoval et al., 2000). This can be known through validity existence facts. The validity existence of an instrument can be identified through content analysis and empirical analysis from instrument score of item response data (Lissitz et al., 2007b).

The criteria of validity are divided into two, namely the predictive validity and concurrent validity. Fernandes (1984) writes that the validity based on criteria is intended to answer the question about the extent to which an instrument can predict the participants' ability in the future (predictive validity) or estimate the ability of other measuring devices in almost the same deadline (concurrent validity). A similar opinion is also expressed by Wright et al. (1996) who says that the instrument is said to have predictive validity if it is able to predict capability in the future. In the analysis of the predictive validity, performances to be predicted are called criteria. The size of the estimated

predictive validity value of an instrument is described by the correlation coefficient between the predictors of those criteria.


The content validity of an instrument is the extent to which the items in the instrument represents the components in the over- all area of the contents of the object to be measured and the extent to which the items reflects behavioral traits that will be measured (Nunnally, 1978; Retnawati, 2016). Meanwhile, Lawrence (1994) explains that content validity is the questionable representation of special abilities that must be measured. Based on this opinion, it can be concluded that the content validity is related to the rational analysis of the domain to be measured to determine the representation of the instrument with the ability to be measured. Construct validity is the validity which shows to what extent the instruments reveal the ability or particular theoretical construct to be measured (Nunnally, 1978; Retnawati, 2016). A construct validation procedure starts from an identification and restriction regarding the variables to be measured and is expressed in terms of a logical construct based on the theory of those variables. From this theory, a practical consequence of the results of measurements on certain conditions is drawn, and this consequence will be tested. If the result is in line with expectations, the instrument is considered to have good construct validity.

Validity is an indispensable term required in an instrument's development. According to Sireci supported by Lissitz et al. (2007a), the validation of instruments used in education should involve the content analysis and empirical analysis of the scores obtained from the instrument and the respondents' response to the items. Content analysis of an instrument is associated with content analysis that later, also needs an empirical analysis to prove the construct validity. Both of these analyses are intended to make instruments in the world of education qualified as a standard measurement instrument.

Method

Before writing a test script, there are certain conditions that need to be considered. Yanto et al. (2019) explains that the quality of the questions is largely determined by the material, construct and language aspects. Each aspect is described in a grid for writing questions which is also used as an instrument for assessing the suitability of the test. Based on the instrument and question bank grid, the question script is then written. An example of a question item that was developed can be seen in the following image.

When Yasmin went to the rice fields she saw a population of storks looking for food. A group of cranes is said to be a population because it has the following characteristics. Except...



a. the morphological form is the same
 b. their physiological functions are the same
 c. can produce fertile offspring
 d. can interbreed
 e. consisting of similar individuals

reason

a. when mating occurs, sterile offspring will be produced
 a. storks have different heights and body sizes depending on age
 b. physiological functions differ depending on age
 c. individuals of different species cannot mate
 d. individuals of the same species can mate and produce fertile offspring
 e. individuals of the same species can mate and produce fertile offspring

Figure 1. Examples of critical thinking question

Validation of questions by experts is an important stage in developing evaluation instruments. In this research, question validation was carried out by material experts and media experts to assess the suitability of the

product. This stage involves an evaluation and revision process based on input from experts. After that, the test instrument that has been validated by experts is then carried out empirical validation, namely testing it on students to determine the validity of the test items.

Table 1. Criteria for Measuring Content Validity

Construction	Relevance	Clarity
1 = poor	1 = not relevant	1 = not clear
2 = fair	2 = item need revision	2 = item need revision
3 = average	3 = item need some revision	3 = item need some revision
4 = good	4 = item relevant but need minor revision	4 = clear but need minor revision
5 = very good	5 = very relevant	5 = very clear

Note: the criteria for measuring content validity based on (Pada et al., 2015)

In the context of this research, validation of questions by experts was carried out to ensure that the questions presented in the evaluation instrument were of good quality and relevant to the material being tested. Data collection was carried out using a Likert type assessment scale validation sheet to measure aspects of each item in the content requested by the Expert. Each Expert is given a Validation Sheet containing 40 Questions and 3 Aspects which are assessed based on construction, relevance and clarity for each test item on a five-point scale as in the following table (Table I).

Table 2. Validation form by Experts

Item	Aspects Assessed														
	Construction					Relevance					Clarity				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Question 1															
Question 2															
Question 3															
Question 4															
Question 5															
Question 6															
Question 7															
Question 8															
...															
Question 40															
Total Score															
Percentage (%)															

Instructions: Please read each element and its benchmarks. Assess the extent of the construction of these elements using the assessment scale provided. Be sure to provide one assessment from the checklist for each element.

To see consistency between validators, the content validity index is calculated using the content validity index to ensure whether the content of the questions is appropriate and relevant to the objectives of the research

itself or not. Content validation can be seen from the test grid. The instrument feasibility validation sheet uses 5 interval scales and validation data analysis uses V Aiken. Data analysis is carried out by: Tabulating validation results by experts and calculating the content validity coefficient using the V Aiken as shown by Formula 1 (Aiken, 1987).

$$V = \frac{\sum S_r}{n(c-1)} \tag{1}$$

Note:

S : r-lo

V: index of expert agreement regarding item validity

R: expert choice category score

Lo: the lowest score in the scoring category

n : number of experts

m : number of items

c : the number of categories that experts can choose from

The Aiken V Index for each item is converted into qualitative data in the range 1 to 0. The results of the validation analysis are compared with the Right-Tail Probabilities (P) for Selected Value of Validity Coefficient (V) table "that for a 5 category scale with 3 validators, the instrument is said to be valid if the coefficient $V = 0.667$ (Azwar, 2012).

Result and Discussion

The content validity test was carried out to determine the extent to which the items in the questions represent all aspects of biological critical thinking skills. Content validity is carried out by looking at the results of the assessments of the experts involved. The content validator assesses the BiCriThiS question items using the validation sheet questionnaire in Appendix 5. The validator as expert judgment assesses the suitability of the question items based on the construction, relevance and clarity of the questions. The assessment results are then analyzed using the Aiken's V formula to determine the Aiken's item index. Interpretation of the validity of question items is determined by valid criteria if the validity index is greater than or equal to 0.667 and invalid if the validity index is less than 0.667 (Azwar, 2012).

At the analysis stage of the content validation test data results carried out by experts, the main focus lies on the construction aspect of the questionnaire. In this case, the construct aspect refers to the clarity and consistency of the questions asked in the questionnaire. The results of content validity measurements show that most of the questions in the questionnaire received positive assessments from experts. Questions number 5, 7, 10, 13, 15, 18, 23, 25, 27, 29, 30, 31, 32, 35, and 38, for example, have high validity values, indicating that the construct aspects of these questions are considered good and in accordance with the research objectives. On the other hand, some questions such as number 20 obtained lower validity scores, which indicates the need to revise and clarify the construct aspects of these questions.

Although overall, the questionnaire received an average validity value of 0.84 which is categorized as

high, it should be noted that several questions received a moderate rating, such as questions number 1, 4, 6, 8, 9, 12, 14, 16, 17, 19, 21, 24, 28, 33, 34, 37, 39, and 40. Therefore, it is necessary to carry out further evaluation of the construct aspects of these questions to ensure that they can measure the variables in question clearly and consistently.

Table 3. Coefficient index of Content Validity for All Aspects

Item	Aspects					
	Construction		Relevance		Clarity	
	V	Note	V	Note	V	Note
Question 1	0.75	Valid	0.83	Valid	0.75	Valid
Question 2	0.92	Valid	1	Valid	0.75	Valid
Question 3	0.92	Valid	0.83	Valid	1	Valid
Question 4	0.75	Valid	1	Valid	1	Valid
Question 5	1	Valid	0.83	Valid	0.83	Valid
Question 6	0.75	Valid	0.75	Valid	0.75	Valid
Question 7	1	Valid	1	Valid	1	Valid
Question 8	0.67	Valid	0.75	Valid	0.75	Valid
Question 9	0.75	Valid	1	Valid	0.75	Valid
Question 10	1	Valid	1	Valid	0.75	Valid
Question 11	0.83	Valid	0.75	Valid	0.75	Valid
Question 12	0.75	Valid	0.75	Valid	0.75	Valid
Question 13	1	Valid	1	Valid	0.75	Valid
Question 14	0.75	Valid	0.83	Valid	0.75	Valid
Question 15	0.83	Valid	0.75	Valid	0.83	Valid
Question 16	0.75	Valid	1	Valid	0.75	Valid
Question 17	0.75	Valid	0.75	Valid	0.75	Valid
Question 18	1	Valid	1	Valid	1	Valid
Question 19	0.75	Valid	0.75	Valid	1	Valid
Question 20	0.50*	Valid	0.92	Valid	1	Valid
Question 21	0.75	Valid	0.67	Valid	1	Valid
Question 22	0.83	Valid	0.92	Valid	1	Valid
Question 23	1	Valid	0.75	Valid	0.75	Valid
Question 24	0.75	Valid	0.75	Valid	1	Valid
Question 25	1	Valid	1	Valid	0.75	Valid
Question 26	0.83	Valid	0.75	Valid	1	Valid
Question 27	1	Valid	0.75	Valid	0.83	Valid
Question 28	0.75	Valid	1	Valid	1	Valid
Question 29	1	Valid	1	Valid	0.83	Valid
Question 30	0.83	Valid	0.83	Valid	0.83	Valid
Question 31	0.83	Valid	0.92	Valid	0.83	Valid
Question 32	0.83	Valid	0.75	Valid	0.83	Valid
Question 33	0.75	Valid	1	Valid	0.75	Valid
Question 34	0.75	Valid	0.75	Valid	0.75	Valid
Question 35	1	Valid	0.83	Valid	0.67	Valid
Question 36	1	Valid	0.75	Valid	1	Valid
Question 37	0.75	Valid	0.83	Valid	1	Valid
Question 38	1	Valid	0.92	Valid	1	Valid
Question 39	0.75	Valid	0.92	Valid	0.75	Valid
Question 40	0.75	Valid	0.67	Valid	0.83	Valid

The results of this content validation provide a strong basis for understanding the quality of the questionnaire and prove that the measurement tool used has high validity. However, improvements to questions with moderate validity values can increase the clarity

and consistency of the overall questionnaire construct. This process will ensure that the questionnaire is reliable and effective in collecting accurate data in accordance with the research objectives.

In the results of content validation test data carried out by experts, the aspect of relevance to the data shows results that can generally be considered high. This can be seen from the average overall validation score of 0.86, which was obtained from the assessment of each question number by experts. Questions with numbers 2, 4, 7, 9, 10, 13, 16, 18, 20, 25, 28, 29, 30, 31, 33, 35, 37, 38, and 39 get a high rating with a value of 1, 00, indicating the optimal level of relevance to the data presented. Although there are several questions with moderate ratings such as questions number 6, 8, 11, 12, 15, 17, 19, 21, 23, 24, 26, 27, 32, 34, 36, and 40, overall, the data shows a trend experts to provide a high assessment of the relevance of the content in the instruments prepared.

Although in general the level of validation of the relevance of the content in the instrument is relatively high, several questions received a moderate rating. This could be a point of attention for further revision or clarification in the preparation of the instrument. For example, questions number 6, 8, 11, and 40 need further attention in order to increase their level of relevance to the data presented. Therefore, suggestions for improvement or clarification from experts regarding aspects that are assessed as moderate need to be considered to ensure that the instrument prepared truly reaches the expected level of relevance.

Overall, the results of the content validation test show that the instrument prepared has a good level of relevance to the data presented. Although some scoring is taking place on some question numbers, the average relevance level of the instrument reaches 0.86, which can be considered high. Therefore, further understanding of suggestions for improvement and clarification from experts can improve the quality of the instrument to ensure the continuation of an optimal level of relevant.

Discussion of the results of content validation test data by experts based on the clarity aspect can be seen from the overall average value and analysis of each question number. In this case, the clarity aspect is measured by the value given by the expert to each question, in the categories high, medium and low.

From the data results, it can be concluded that the majority of questions received high marks, especially questions with numbers 3, 4, 5, 7, 18, 19, 20, 21, 22, 24, 26, 28, 29, 30, 31, and 32, all of which have a maximum value of 1.00. This shows that the expert gave a very good assessment of the clarity of these questions. On the other hand, there are several questions with medium ratings, namely questions numbered 1, 2, 6, 8, 9, 10, 11, 12, 13, 14, 16, 17, 23, 25, 33, 34, and 39. However, there were no questions that received low marks.

The overall average score of 0.85 indicates that the clarity aspect of this research instrument can be categorized as high. Thus, it can be concluded that experts consider the contents of this research instrument to be generally easy to understand, clear and not cause confusion. Even though there are several questions with moderate ratings, the high average score indicates that this instrument as a whole meets the criteria for good clarity. Therefore, the results of content validation from the clarity aspect can be considered satisfactory.

Thus, the conclusion from the results of the discussion is that this biological critical thinking skills evaluation instrument has passed the content validity test well. All questions were assessed as valid and relevant, although some minor corrections may be required to improve certain aspects. Overall, the expert evaluation provides a positive picture of the quality of this research instrument, providing a strong basis for its use in the context of measuring critical thinking skills in the field of biology.

Conclusion

The conclusion from the results of the discussion is that this biological critical thinking skills evaluation instrument has passed the content validity test well. All questions were assessed as valid and relevant, although some minor corrections may be required to improve certain aspects. Overall, the expert evaluation provides a positive picture of the quality of this research instrument, providing a strong basis for its use in the context of measuring critical thinking skills in the field of biology.

Acknowledgments

Irma Kharisma is the part of the Biology Education, Faculty of Teacher Training and Education, Syiah Kuala University, prepared this journal article based on the report The Development of Computerized Based Test Instrument for Critical Thinking Skill on Ecology Lesson for High School Education Level. The opinions expressed here in are those of the authors and do not necessarily reflect the views of funding agency.

Author Contributions

Conceptualization, F. S, H. R, W. A; methodology, F. S; software, F. S; validation, H. R, W. A, I. H, S; formal analysis, F. S, H. R, W. A; investigation, F. S; resources, F. S; data curation, F. S, H. R, W. A; writing—original draft preparation, F. S; writing—review and editing, F. S, H. R, W. A; visualization, F. S; supervision, I. H, S; project administration, F. S; funding acquisition, F. S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

References

- Aiken, L. R. (1987). Formulas for Equating Ratings on Different Scales. *Educational and Psychological Measurement*, 47(1), 51–54. <https://doi.org/10.1177/0013164487471007>
- Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.
- Azwar, S. (2012). *Reliabilitas dan validitas edisi 4*. Pustaka Pelajar.
- Cullinane, A., & Liston, M. (2011). *Two-tier Multiple Choice Question: An Alternative Method of Formatif Assessment for First Year Undergraduate Biology Students*. National Center for Excellence In Mathematics and Education Science Teaching and Learning (NCE-MSTL).
- Elliot, D. L., & Kobayashi, S. (2019). How can PhD supervisors play a role in bridging academic cultures? *Teaching in Higher Education*, 24(8), 911–929. <https://doi.org/10.1080/13562517.2018.1517305>
- Ennis, R. (2011). Critical Thinking. *Inquiry: Critical Thinking Across the Disciplines*, 26(2), 5–19. <https://doi.org/10.5840/inquiryctnews201126215>
- Ennis, R. H. (n.d.). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, 18(3), 4–10. <https://doi.org/10.3102/0013189x018003004>
- Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. In *Educational Leadership* (Vol. 43, Issue 2, pp. 44–48).
- Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice*, 32(3), 179–186. <https://doi.org/10.1080/00405849309543594>
- Gillies, R. M., Nichols, K., Burgh, G., & Haynes, M. (2012). The effects of two strategic and meta-cognitive questioning approaches on children's explanatory behaviour, problem-solving, and learning during cooperative, inquiry-based science. *International Journal of Educational Research*, 53, 93–106. <https://doi.org/10.1016/j.ijer.2012.02.003>
- Hidayah, N., Yuliana, A. T., & Hanafi, H. (2020). Theoretical Validity of Problem Focused-Coping Skill Guideline to Develop Students' Critical Thinking Skills. *Jurnal Kajian Bimbingan Dan Konseling*, 5(4), 183–191. <https://doi.org/10.17977/um001v5i42020p183>
- Istiyono, E. (2014). Pengukuran kemampuan berpikir tingkat tinggi fisika peserta didik SMA di DIY. *Jurnal Sains Dan Teknologi*, 5, 1. Retrieved from <https://eprints.uny.ac.id/13115/>
- Istiyono, E., Dwandaru, W. S. B., Ledo, Y. A., Rahayu, F., & Nadapdap, A. (2019). Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge. *International Journal of Instruction*, 12(4), 267–280. <https://doi.org/10.29333/iji.2019.12417a>
- Istiyono, E., Dwandaru, W. S. B., Permatasari, A. K., & Ariatiawan. (2020). Developing computer based test to assess students' problem-solving in physics learning. *Journal of Physics: Conference Series*, 1440(1), 012060. <https://doi.org/10.1088/1742-6596/1440/1/012060>
- Istiyono, E., 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>
- Johnson, D. W., Johnson, R. T., & Smith, K. A. (2014). Cooperative Learning: Improving University Instruction by Basing Practice on Validated Theory. *Journal of Excellence in College Teaching*, 25(4), 85–118. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10180297>
- Kennedy, M., Fisher, M., & Ennis, R. (2013). Critical Thinking: Literature Review and Needed Research. In *Educational Values and Cognitive Instruction: Implications for Reform* (Vol. 2, pp. 11–40).
- Linn, R. L., & Gronlund, N. E. (1995). *Measurement and evaluation in teaching*. Macmillan.
- Lissitz, R. W., & Samuelsen, K. (2007a). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, 36(8), 437–448. <https://doi.org/10.3102/0013189x07311286>
- Lissitz, R. W., & Samuelsen, K. (2007b). Further Clarification Regarding Validity and Education. *Educational Researcher*, 36(8), 482–484. <https://doi.org/10.3102/0013189x07311612>
- Mundilarto. (2002). *Kapita Selektta Pendidikan Fisika*. FMIPA UNY.
- Nunnally, J. C. (1978). An Overview of Psychological Measurement. In *Clinical Diagnosis of Mental Disorders* (pp. 97–146). Springer US. https://doi.org/10.1007/978-1-4684-2490-4_4
- Pada, A. U. T., Kartowagiran, B., & Subali, B. (2015). Content Validity of Creative Thinking Skills Assessment. *Proceeding of International Conference On Research, Implementation And Education Of Mathematics And Sciences, May*, 17–19. Retrieved from <https://core.ac.uk/download/pdf/33519344.pdf>
- Pozo-Muñoz, C., Reboloso-Pacheco, E., & Fernández-Ramírez, B. (2000). The "Ideal Teacher".

- Implications for student evaluation of teacher effectiveness. *Assessment & Evaluation in Higher Education*, 25(3), 253-263. <https://doi.org/10.1080/02602930050135121>
- Putri, F. S., Istiyono, E., & Nurcahyanto, E. (2016). Pengembangan Instrumen Tes Keterampilan Berpikir Kritis Dalam Bentuk Pilihan Ganda Beralasan (Politomus) Di DIY. *Unnes Physics Education Journal*, 5(2). <https://doi.org/10.15294/upej.v5i2.13626>
- Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *REID (Research and Evaluation in Education)*, 2(2), 155-164. <https://doi.org/10.21831/reid.v2i2.11029>
- Ritonga, S., Safrida, S., Huda, I., Supriatno, & Sarong, M. A. (2020). The effect of problem-based video animation instructions to improve students' critical thinking skills. *Journal of Physics: Conference Series*, 1460(1), 012107. <https://doi.org/10.1088/1742-6596/1460/1/012107>
- Saleh, S. E. (2019). Critical Thinking as a 21 St Century Skill: Conceptions, Implementation, and Challenges in the EFL Classroom. *European Journal of Foreign Language Teaching*, 4(1), 1-16. <https://doi.org/10.46827/ejfl.v0i0.2209>
- Sandoval, A. M. R., Hancock, D., Poythress, N., Edens, J. F., & Lilienfeld, S. (2000). Construct validity of the psychopathic personality inventory in a correctional sample. *Journal of Personality Assessment*, 74(2), 262-281. https://doi.org/10.1207/S15327752JPA7402_7
- Shaaban, K. A. (2014). Assessment of critical thinking skills through reading comprehension. *International Journal of Language Studies*, 8(2), 117-140. Retrieved from <http://proxy.libraries.smu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cms&AN=95416342&site=ehost-live&scope=site>
- Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*, 17(4), 396-408. <https://doi.org/10.1037/1040-3590.17.4.396>
- Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development* (Issue March). STORRE: Stirling Online Research Repository. Retrieved from <https://www.storre.stir.ac.uk/handle/1893/32425>
- Syaifuddin, M., Darmayanti, R., & Rizki, N. (2022). Development Of A Two-Tier Multiple-Choice (TTMC) Diagnostic Test For Geometry Materials To Identify Misconceptions Of Middle School Students. *Jurnal Silogisme : Kajian Ilmu Matematika Dan Pembelajarannya*, 7(2). <https://doi.org/10.24269/silogisme.v7i2.5456>
- Winarti, Cari, Suparmi, Sunarno, W., & Istiyono, E. (2017). Development of two tier test to assess conceptual understanding in heat and temperature. *Journal of Physics: Conference Series*, 795(1), 12052. <https://doi.org/10.1088/1742-6596/795/1/012052>
- Wright, G., Lawrence, M. J., & Collopy, F. (1996). The role and validity of judgment in forecasting. *International Journal of Forecasting*, 12(1), 1-8. [https://doi.org/10.1016/0169-2070\(96\)00674-7](https://doi.org/10.1016/0169-2070(96)00674-7)
- Yanto, B. E., Subali, B., & Suyanto, S. (2019). Improving Students' Scientific Reasoning Skills through the Three Levels of Inquiry. *International Journal of Instruction*, 12(4), 689-704. <https://doi.org/10.29333/iji.2019.12444a>