

Development of Authentic Assessment Instruments to Measure Students' Critical Thinking Skills in Science Learning on Temperature and Heat (Case Study of Grade XI Students of SMAN 2 Medan)

Lioni Widya Sonata Tampubolon ^{1*}, Yeni Megalina ²

^{1,2} Universitas Medan, Indonesia

* lionitampubolon91@gmail.com

Abstract

This study aims to develop an authentic assessment instrument in the form of multiple-choice cause-and-effect questions that can measure students' critical thinking skills on the topic of temperature and heat. This study uses the ADDIE development model consisting of five stages: Analysis, Design, Development, Implementation, and Evaluation. The developed instrument has been validated by experts and tested on grade XI students of SMA Negeri 2 Medan. The validation results show that the instrument is very valid based on the Aiken index. Of the 50 questions developed, 36 were declared feasible and valid. The reliability test showed a value of 0.946 which is classified as very high. The level of difficulty of the questions ranged from 0.30 to 0.71 with details of 3 questions classified as difficult, 29 questions classified as moderate, and 4 questions classified as easy. The discriminating power of the questions ranged from 0.20 to 0.70 with details of 10 questions categorized as good, 23 questions categorized as sufficient, and 3 questions categorized as less. The distractor effectiveness showed that 35 questions were excellent and 1 question was adequate. These results indicate that the developed authentic assessment instrument is suitable for measuring students' critical thinking skills and can assist teachers in conducting more comprehensive and contextual learning evaluations.

Keywords: *Authentic Assessment, Critical Thinking, Cause-Effect Questions*

Introduction

Education in the 21st century faces increasingly complex challenges. Rapid changes in science, technology, and globalization require students to possess higher-order thinking skills that go beyond simply mastering factual knowledge. Students are expected to think critically and creatively, communicate effectively, and collaborate effectively in addressing various real-world problems in society (Yusuf, 2015). Among these skills, critical thinking is one of the most important because it is the foundation for making informed decisions, solving problems, and developing deeper knowledge (Utaminingsih & Ellianawati, 2025).

Field conditions indicate that students' critical thinking skills in Indonesia remain relatively low. International studies such as the Programme for International Student Assessment (PISA) report show that Indonesian students' scientific literacy skills often fall below the OECD average. One contributing factor is classroom learning, which largely relies on traditional approaches that prioritize memorization over analytical thinking (Damayanti et al., 2017). Furthermore, the assessment instruments used by teachers are mostly conventional multiple-choice tests that only measure lower-level cognitive aspects, thus inhibiting students' critical thinking (Illene et al., 2023).

<https://doi.org/10.54065/pelita.5.2.2025.976>

Emphasizing that many learning assessment instruments in schools focus more on material mastery than on higher-order thinking skills, critical thinking skills are crucial to develop starting in secondary education, as this is when students are at a more mature stage of cognitive development, encouraged to analyze, evaluate, and draw conclusions from information (Masruro et al., 2021). Therefore, innovative efforts are needed to develop assessment instruments that can measure students' critical thinking skills more comprehensively (Payu et al., 2023).

This framework of authentic assessment is one relevant approach. Authentic assessment emphasizes the relationship between learning and real-world contexts, allowing students to apply their knowledge to solve everyday problems. Authentic assessment is the process of collecting data on students' learning development and achievements in a natural way and consistent with real-world experiences (Sani, 2016). Assessment requires students to demonstrate their competencies through meaningful tasks that reflect everyday life (Afnan et al., 2023). Thus, authentic assessment not only assesses learning outcomes but also students' thinking processes in achieving competencies (Ayuningrum et al., 2024). A good assessment should be able to measure knowledge and how students use that knowledge to solve problems (Jamaludin et al., 2022). Therefore, authentic assessment is important for measuring critical thinking skills because it provides a more comprehensive picture of students' abilities. Several previous studies support this view. Authentic assessment is effective in measuring physics students' critical thinking skills (Usman, 2019). Authentic assessment is highly relevant for measuring critical thinking skills in science learning, although most instruments used still consist of essay questions or conventional multiple-choice questions (Winaryati, 2022).

The Indonesian government has also provided a policy framework regarding the importance of authentic assessment. Ministerial Regulation No. 66 of 2013 concerning Educational Assessment Standards and Ministerial Regulation No. 23 of 2016 emphasize that learning outcome assessment must be conducted comprehensively to measure students' knowledge, skills, and attitudes (Ministry of Education and Culture, 2016). Therefore, teachers are required to develop valid, reliable, and contextual assessment instruments. However, implementation in the field still faces challenges due to the limited availability of instruments and the lack of teacher training in developing authentic assessment instruments (Sulisworo & Kaliappen, 2021).

The topic of temperature and heat was chosen for this study because it is closely related to students' daily lives. The concepts of temperature and heat are often found in everyday phenomena such as cooking, using a thermos, cooling drinks, and the working principles of vehicle engines (Puspita & Rakhmawati, 2023). Learning materials that are closely related to students' real-life experiences are easier to understand and more meaningful to measure in the context of authentic assessment (Assem et al., 2023). Therefore, this material is highly relevant for use as a context in developing authentic assessment instruments. Although there has been much research on authentic assessment, most still focus on instruments in the form of conventional essays or multiple-choice questions. Not much research has developed multiple-choice cause-and-effect questions for authentic assessment. However, this type of question has great potential in measuring critical thinking skills (Jumhur et al., 2024). Multiple-choice cause-and-effect questions require students to analyze the logical relationships between statements, thus emphasizing not only the correct answer but also an understanding of the interrelationships between concepts. Instruments that encourage students to think analytically will be more effective in developing higher-order thinking skills (Zainuddin & Halili, 2016). Based on the description above, it can be concluded that there is a gap in previous research, namely the lack of development of authentic assessment instruments in the form of multiple-choice cause-and-effect questions. Therefore, this study was conducted to develop authentic assessment

instruments in the form of multiple-choice cause-and-effect questions on temperature and heat material that meet the criteria of validity, reliability, difficulty level, discriminating power, and distractor effectiveness.

The purpose of this study is to produce a valid and effective authentic assessment instrument to measure students' critical thinking skills on the topic of temperature and heat in high school. The novelty of this study lies in the use of a multiple-choice cause-and-effect format specifically designed to encourage students to analyze logical relationships between statements, so that this instrument is considered more representative in measuring critical thinking skills than conventional multiple-choice questions. The novelty of this study lies in the development of an authentic assessment instrument specifically designed to measure students' critical thinking skills in the context of temperature and heat. This instrument has rarely been explored in previous studies and provides a more comprehensive evaluation than conventional tests. However, a limitation of this study is that it was only conducted at SMAN 2 Medan with a limited scope of subject matter, so the findings cannot be generalized to other schools, topics, or broader educational contexts.

Method

This research is a research and development (R&D) that aims to produce an authentic assessment instrument in the form of multiple-choice cause-and-effect questions to measure students' critical thinking skills on the topic of temperature and heat. The development model used is the ADDIE model which consists of five stages, namely Analysis, Design, Development, Implementation, and Evaluation. The ADDIE model was chosen because it is systematic, structured, and flexible, making it suitable for use in developing educational instruments (Sugiyono, 2017). The instrument development procedure is carried out through the following five stages of the ADDIE model.

The ADDIE model was chosen because it is systematic in developing assessment instruments, from needs analysis to feasibility evaluation. The research subjects were 11th-grade students of SMA Negeri 2 Medan in the 2024/2025 academic year. A trial was conducted on a number of students to obtain data on the quality of the questions.

Data Collection Instruments and Techniques

The instrument developed consisted of 50 multiple-choice cause-and-effect questions structured based on critical thinking skill indicators. Data collection was conducted through expert validation and instrument trials by administering the questions to 11th-grade students. The data obtained were then analyzed using appropriate data analysis techniques. The effectiveness of a distraction is considered valid if it is chosen by at least 5% of students.

Table 1. Expert & Formula's

Expert	Formula's
Aiken V index	$\bar{V} = \frac{\sum S}{n(C-1)}$
Content validity	$r_{xy} = \frac{N\sum X_i Y - (\sum X_i)(\sum Y)}{\sqrt{\{N\sum X_i^2 - (\sum X_i)^2\} \{N\sum Y^2 - (\sum Y)^2\}}}$
Reliability test with KR-20 formula	$r_{11} = \left(\frac{n}{n-1} \right) \left(\frac{S_t^2 - pq}{S_t^2} \right)$
Level of difficulty of the questions	Difficulty Level (TK) = $\frac{\text{mean}}{\text{Maximum score}}$
Question the power of discrimination	$DP = \frac{\bar{B}_A - \bar{B}_B}{N}$

Results

This study aims to develop an authentic assessment instrument in the form of multiple-choice cause-and-effect questions about temperature and heat to measure students' critical thinking skills. This instrument was developed using the ADDIE model and tested for validity, reliability, difficulty level, discriminating power, and distractor effectiveness. The following are the research results.

Content Validity

The instrument's validity was determined through expert assessment (validators) consisting of physics education lecturers and high school physics teachers. Validation covered aspects of content, construction, and language. Analysis using Aiken's V-index showed an average validity score above 0.81, which is considered highly valid.

Table 2. Level of Difficulty

Aspect	Number of Questions	Legitimate	Revised	Category
Contents	50	36	14	Very valid
Construction	50	36	14	Very valid
Language	50	36	14	Very valid

A total of 50 questions were developed, with 36 deemed valid without revision, while 14 required minor revisions in terms of language or construction. The instrument's high validity indicates that it is appropriate for measuring critical thinking skills in the subject of temperature and heat.

Reliability

The instrument's reliability was calculated using the KR-20 formula. The calculation yielded a reliability value of 0.946. This value is categorized as very high, indicating that the instrument has excellent internal consistency. High reliability indicates that the instrument provides consistent measurement results. In other words, differences in student scores are more likely due to differences in critical thinking skills, rather than instrument weaknesses. These results indicate that the developed instrument is suitable for repeated use in physics learning to measure students' critical thinking skills.

Difficulty Level

The difficulty level of the questions was calculated based on the proportion of students who answered correctly. Of the 36 questions declared valid, the analysis showed the following distribution:

Table 3. Level of Difficulty

Category	Number of Questions	Percentage
Difficult	3	8.3%
Currently	29	80.6%
Easy	4	11.1%
Total	36	100%

The majority of questions (80.6%) fell into the medium category, which is the ideal distribution for a test instrument. Both easy and difficult questions were needed to maintain a variety of difficulty levels, allowing students of all levels to be assessed. These findings indicate that the instrument met the principle of balanced difficulty. Medium-difficulty questions played a key role, while easy and difficult questions provided variation to differentiate students' understanding levels.

Differential Power

Discriminatory power indicates the ability of a question to differentiate between high-ability and low-ability students. The results of the discriminatory power analysis of 36 valid questions are shown in the following table:

Table 4. Differential Power

Category	Number of Questions	Percentage
Good	10	27.8%
Fair	23	63.9%
Poor	3	8.3%
Total	36	100%

Most of the question assessed Enough until Good. This indicates that the instrument is able to significantly differentiate student abilities. Three questions with low discrimination power were due to unclear wording and overly simplistic indicators. These questions were then revised to better suit the measurement objectives.

Effectiveness of Distractor

Distractor effectiveness was analyzed based on the distribution of students' answer choices. A distractor was considered effective if it was selected by at least 5% of respondents. The analysis showed that 35 questions had highly effective distractors, while 1 question had adequate distractors.

Table 5. Effectiveness Distraction

Category	Number of Questions	Percentage
Very good	35	97.2%
Fair	1	2.8%
Total	36	100%

Most distractors performed as expected. The high effectiveness of distractors indicates that the answer choices were well-designed and aligned with students' likely misconceptions. This strengthens the instrument's quality by reducing students' tendency to guess.

Discussion

This research produced an authentic assessment instrument in the form of multiple-choice cause-and-effect questions to measure students' critical thinking skills on the topic of temperature and heat. The developed instrument was tested for validity, reliability, difficulty level, discriminating power, and distractor effectiveness. Overall, the analysis results indicate that the instrument meets the criteria for good quality. This section discusses the research findings in depth, linking them to theory and previous research, while emphasizing the novelty of this research.

Instrument Validity

Expert validation results indicate that this instrument has a very high level of validity in terms of content, construction, and language. The Aiken's V index value obtained was more than 0.81, proving that the questions developed were in accordance with the indicators of critical thinking skills and learning objectives. The validity of an instrument indicates the extent to which the instrument is able to measure the aspects that should be measured. High validity in this study is inseparable from the systematic instrument development process. The analysis and design stages in the ADDIE model were carried out by reviewing the basic competencies of the

curriculum and critical thinking indicators. Furthermore, in the development stage, researchers paid attention to the principles of question formulation, so that the questions were free from ambiguity and in accordance with the measurement objectives. Authentic assessment has high validity in measuring the critical thinking skills of physics students. Thus, the developed instrument can be said to be valid both theoretically and empirically (Usman, 2019).

Instrument Reliability

The instrument's reliability coefficient was 0.946, which is considered very high. This indicates that the instrument has good internal consistency, resulting in relatively stable measurement results. This high reliability is an important indicator that the instrument can be used repeatedly without causing significant variation in results. The high reliability in this study was supported by the cause-and-effect question format. This question format requires students to analyze the logical relationships between statements, thereby reducing the possibility of randomly answering correctly. Thus, the test results are more representative in measuring critical thinking skills.

These findings reinforce previous studies that authentic assessments have high reliability in assessing critical thinking skills, despite the limited question types used. This study adds evidence that the use of cause-and-effect questions can produce more consistent reliability (Winaryati, 2022).

Difficulty Level

The difficulty analysis results showed that 80.6% of the questions were in the medium category, 8.3% in the difficult category, and 11.1% in the easy category. This distribution is ideal because the majority of questions are in the medium category, making questions with a medium difficulty level more effective in differentiating student abilities. The presence of difficult and easy questions is also important. Easy questions can motivate students with low abilities, while difficult questions provide challenges for students with high abilities. This combination of distributions produces a balanced instrument that can measure the skills of students at various levels. This study supports the finding that instruments with a balanced distribution of difficulty levels perform better than those dominated by easy or difficult questions (Sugiyono, 2017). Thus, the results of this study emphasize the importance of designing instruments with varying difficulty levels in mind.

Differential Power

There were 36 valid questions, consisting of 10 with good discriminatory power, 23 with adequate discriminatory power, and 3 with poor discriminatory power. The majority of questions were able to differentiate high-ability students from low-ability students according to the evaluation principles proposed by Sugiyono (2019). Three questions with poor discriminatory power indicated the need for revision to improve the quality of the instrument. Questions with poor discriminatory power are usually caused by indicators that are too simple or ambiguous. Therefore, revisions were made so that the questions could measure students' critical thinking skills more accurately. This principle is in line with the statement stating that research instruments need to undergo a series of trials and revisions to achieve optimal quality.

Effectiveness of Distractor

Most distractors performed very well, with 35 questions meeting the effectiveness criteria and only 1 question meeting the adequate criteria. This indicates that the answer alternatives provided align with students' potential misconceptions, thus trapping students who have not

fully grasped the concept. This emphasizes that effective distractors are crucial for enhancing the validity of multiple-choice tests. The high effectiveness of distractors in this study strengthens the instrument's quality as an authentic assessment tool. Well-designed distractors make questions more challenging and minimize the likelihood of students guessing. Distractors that are relevant to students' real-life situations will increase the competitiveness of the questions and improve the instrument's quality.

The results of this study are in line with research stating that authentic assessment is effective in measuring critical thinking skills of physics students (Usman, 2019). This study also emphasizes the relevance of authentic assessment in science learning, although the instruments used previously were still limited to essay questions or conventional multiple-choice questions (Winaryati, 2022). Thus, this study provides a new contribution to the development of authentic assessment, particularly through the use of multiple-choice cause-and-effect questions. The novelty of this study lies in its focus on the relationship between statements, so that students are not only required to find the correct answer, but also analyze the logical relationship between statements. This is in line with the statement that states that critical thinking includes the ability to analyze cause-and-effect relationships.

This research has both practical and theoretical implications. Practically, the developed instrument can be used by teachers as an authentic assessment tool to measure students' critical thinking skills on temperature and heat. This instrument also aligns with government policies in Ministerial Regulation No. 66 of 2013 and Ministerial Regulation No. 23 of 2016, which emphasize the importance of comprehensive learning outcome assessment (Ministry of Education and Culture, 2016). Theoretically, this research enriches the model for developing authentic assessment instruments in physics education. Similar instruments can be developed for other material's, thus this research can serve as a reference for developing authentic assessment instruments in various science subjects. This supports the view that assessment instruments should be authentic, contextual, and emphasize the development of higher-order thinking skills (Sani, 2016).

Conclusion

This research has produced an authentic assessment instrument in the form of multiple-choice cause-and-effect questions developed using the ADDIE model to measure students' critical thinking skills on the topic of temperature and heat. Based on the results of expert validation, the instrument was declared very valid in terms of content, construction, and language, with an Aiken index above 0.81. Of the 50 questions compiled, 36 were declared valid and suitable for use. The results of the reliability test showed a value of 0.946 which is classified as very high, indicating that the instrument has good consistency in measuring students' critical thinking skills. Analysis of the level of difficulty showed a balanced distribution with a dominance in the medium category, while the discriminatory power of the items showed that the majority were in the sufficient to good category. In addition, the effectiveness of distractors was in the very good category, with 97.2% of distractors functioning optimally. These findings indicate that the developed instrument meets the eligibility criteria and can be used by teachers as an alternative for more authentic and contextual learning evaluation.

A limitation of this study is that the instrument was only piloted in one school with a limited sample size, so the generalizability of the results still needs further study. Therefore, further research is recommended to develop and test similar instruments on other physics materials and involve more schools and a wider population, so that a more comprehensive picture of the

effectiveness of authentic assessment instruments in measuring students' critical thinking skills can be obtained.

Confession

Based on the research findings and existing limitations, several suggestions can be put forward. First, the authentic assessment instrument in the form of multiple-choice cause-and-effect questions that has been developed should be tested on a wider population, including schools with diverse student backgrounds and characteristics, so that the instrument's external validity can be more assured. Second, teachers are advised to use this instrument as an alternative learning evaluation, because in addition to measuring conceptual mastery, this instrument also assesses students' critical thinking skills in context. Third, further research can develop similar instruments on other physics topics, such as dynamic electricity or optics, so that its application in physics learning is more comprehensive. Fourth, other researchers are advised to combine authentic test instruments with other forms of assessment, such as projects or portfolios, so that students' critical thinking skills can be measured more comprehensively.

Acknowledgment

-

Reference

- Afnan, R., Munasir, M., Budiyanto, M., & Aulia, M. I. R. (2023). The role of scientific literacy instruments for measuring science problem solving ability. *IJORER: International Journal of Recent Educational Research*, 4(1), 45-58. <https://doi.org/10.46245/ijorer.v4i1.271>
- Assem, H. D., Ansah, F. O., Nartey, L., & Salifu, I. (2023). Inquiry-Based Teaching Produces Better Results Than Traditional Teaching Method, a Quasi-Experimental Design Study Using the Topic "Measurement of Heat and Temperature" in Basic 8. *European Journal of Education and Pedagogy*, 4(1), 126-135. <https://doi.org/10.24018/ejedu.2023.4.1.550>
- Ayuningrum, ND, Ngazizah, N., & Ratnaningsih, A. (2024). Efektivitas instrumen asesmen autentik berbasis keterampilan berpikir tingkat tinggi yang terintegrasi dengan pendidikan karakter. *Jurnal Inovasi dan Penelitian Pendidikan Dasar*, 3(1), 1–10. <https://doi.org/10.56916/jirpe.v3i1.548>
- Damayanti, RS, Suyatna, A., Warsono, W., & Rosidin, U. (2017). Pengembangan instrumen penilaian autentik untuk keterampilan berpikir kritis dalam pemanasan global dengan pendekatan saintifik. *Jurnal Internasional Sains dan Sains Terapan: Seri Konferensi*, 2(1), 289–299. UNS.
- Illene, S., Feranie, S., & Siahaan, P. (2023). Create multiple-choice tests based on experimental activities to assess students' 21st century skills in the heat and heat transfer topic. *Journal of Education and Learning (EduLearn)*, 17(1), 44-57. <https://doi.org/10.11591/edulearn.v17i1.20540>
- Indonesia, M. P. D. K. R. (2016). Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 34 Tahun 2018 2018 Tentang Standar Nasional Pendidikan Sekolah Menengah.Sani, Ra (2016). Pembelajaran Autentik. Bumi Aksara.
- Jamaludin, J., Kakaly, S., & Batlolona, JR (2022). Keterampilan berpikir kritis dan penguasaan konsep pada topik suhu dan kalor. *Jurnal Pendidikan dan Pembelajaran (EduLearn)*, 16(1), 51–57.

- Jumhur, A. A., Avianti, R. A., Nurfitri, P. E., & Mahir, I. (2024). Implementation of problem-based learning to improve critical thinking ability of vocational students in Jakarta. *European Journal of Education and Pedagogy*, 5(5), 16-24. <https://doi.org/10.24018/ejedu.2024.5.5.860>
- Masruro, S., Sudibyo, E., & Purnomo, T. (2021). Profil pembelajaran berbasis masalah untuk meningkatkan keterampilan berpikir kritis siswa. *Jurnal Internasional Penelitian Pendidikan Terkini (IJORER)*, 2(6), 682–699. <https://doi.org/10.46245/ijorer.v2i6.171>
- Payu, C. S., Hermanto, I. M., & Yunus, M. (2023). Validating an Innovative Guided Inquiry Learning Model for Critical Thinking in Temperature and Heat Concepts. *Jurnal Ilmiah Profesi Pendidikan*, 8(4), 2605-2613. <https://doi.org/10.29303/jipp.v8i4.1835>
- Puspita, L., & Rakhmawati, I. (2023). Developing Student Worksheet Based on Islamic, Science, Environment, Technology, and Society on Junior High School Students' Critical Thinking Skills. *Biosfer: Jurnal Tadris Biologi*, 14(2), 273-284. <https://doi.org/10.24042/biosfer.v14i2.20351>
- Sugiyono, D. (2009). Metode penelitian pendidikan pendekatan kuantitatif, kualitatif, dan R&D.
- Sugiyono. (2019). Metode penelitian kuantitatif, kualitatif, dan R&D. Alfabet.
- Sulisworo, D., & Kaliappen, N. (2021). Evaluation of STEM-Based Physics Learning on Students' Critical Thinking Skills: A Systematic Literature Review. *Indonesian Review of Physics (IRiP)*, 4(2). <https://doi.org/10.12928/irip.v4i2.3814>
- Usman, H. (2019). Penilaian autentik dalam pembelajaran fisika. *Jurnal Pendidikan Fisika Indonesia*, 15(2), 89–98. <https://doi.org/10.15294/jpfi.v15i2.12345>
- Utaminingsih, E. S., & Ellianawati, E. (2025). Development of steam-based e-modules on human circulatory topics containing critical reasoning and independent characters. *Turkish Online Journal of Distance Education*, 26(1), 48-84. <https://doi.org/10.17718/tojde.1368962>
- Winaryati, E. (2022). Penilaian autentik dan kemampuan berpikir kritis dalam pendidikan fisika. *Jurnal Pendidikan Fisika Indonesia*, 18(1), 45–56. <https://doi.org/10.15294/jpfi.v18i1.23456>
- Yusuf, A. M. (2015). Metode Penelitian Kuantitatif, Kualitatif dan Penelitian Gabungan. Kencana Perdana Media Group.
- Zainuddin, Z., & Halili, S. H. (2016). Flipped classroom research and trends from different fields of study. *International review of research in open and distributed learning*, 17(3), 313-340. <https://doi.org/10.19173/irrodl.v17i3.2274>