Berkala Ilmiah Pendidikan Fisika
https://ppjp.ulm.ac.id/journal/index.php/bipf
e-ISSN: 2549-2764 p-ISSN: 2337-604X
Vol 14 No 2 2026
DOI: 10.20527/bipf.v14i2.25819

Development and Validation of a Physics Test Instrument to Measure Problem-Solving Skills on Heat Concepts

Adhelia Tasya Youlanda and Mukhayyarotin Niswati Rodliyatul Jauhariyah*
Physics Education Study Program, Universitas Negeri Surabaya, Surabaya, Indonesia
*mukhayyarotinjauhariyah@unesa.ac.id

Article Info

Article history:
Received: April 5, 2026
Accepted: June 14, 2026
Published: July 25, 2026

Keywords:
Heat
Physics education
Problem-solving skills
Test instrument
Validity

This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

© 2026 Berkala ilmiah Pendidikan Fisika

Abstract

This study aims to develop a valid and reliable knowledge test instrument to measure students' problem-solving skills in heat material. The research employed a research and development approach using the ADDIE development model. The instrument consisted of 10 essay questions constructed based on Polya's problem-solving indicators. The instrument was validated by three experts, specifically two lecturers in the physics department and one physics teacher, and also tested on thirty-three first-year students enrolled in the Physics Education Study Program at Surabaya State University. The analysis included theoretical validity, empirical validity, reliability, difficulty level, and discrimination power. The results showed that the theoretical validity reached an average Percentage of Agreement (PoA) of 88.09% in the valid category. Empirical validity indicated that all items (100%) were valid with a reliability coefficient of 0.83, indicating high reliability. The difficulty level consisted of 60% easy and 40% moderate items, while the discrimination power showed 10% good and 90% sufficient categories. These findings indicate that the developed instrument is suitable for measuring students' problem-solving skills in heat material and supporting the assessment of higher-order thinking skills in physics learning.

To cite this article: Youlanda, A. T. & Jauhariyah, M. N. R. (2026). Development and Validation of a Physics Test Instrument to Measure Problem-Solving Skills on Heat Concepts. *Berkala Ilmiah Pendidikan Fisika, 14*(2), 310-324.

INTRODUCTION

The 21st century is marked by rapid advancements in science and technology that significantly influence many aspects of human life, including the field of education. These developments require education systems to continuously adapt to the demands of globalization and digitalization in order to remain relevant and responsive to societal changes (Hidayati & Julianto, 2025; Lubis et al., 2023). Consequently, education is no longer focused solely on mastering conceptual knowledge but also on developing competencies that are essential for facing modern challenges. Educational institutions are therefore expected to produce human resources who possess strong critical and creative thinking skills and are able to adapt effectively to dynamic and complex changes. Consequently, 21st-century learning requires innovative and learner-centered approaches that encourage students to actively construct knowledge (Adeoye & Jimoh, 2023; Osborne & Allchin, 2025). One of the key goals of this approach is the development of higher-order thinking skills, including critical, creative, collaborative, and communicative abilities needed to solve real-life problems.

Effective learning should be designed to actively involve students in the problem-solving process. In this context, teachers act as facilitators who guide students in

developing systematic and reflective thinking skills. The learning process emphasizes not only obtaining the correct answer but also understanding the problem and determining appropriate solution strategies (Latifah et al., 2023). In science education, higher-order thinking skills are closely connected to problem-solving abilities because scientific learning requires students to analyze phenomena, apply reasoning, and make decisions based on scientific concepts and evidence (Hadi et al., 2023). Therefore, the success of science learning is not only measured by students' ability to memorize concepts but also by their ability to apply these concepts to solve problems and understand scientific phenomena more deeply (Ekawati et al., 2024). However, several international reports reveal that Indonesian students' science competencies are still relatively low, indicating an urgent need to strengthen the development of problem-solving skills, particularly in the physics learning (Bhakti et al., 2025; Brundage et al., 2024a).

One indicator commonly used to evaluate students' scientific abilities at the international level is the Programme for International Student Assessment (PISA), which measures students' capacity to apply scientific knowledge in real-life contexts. According to the 2022 PISA report, Indonesia's science score was 383, which is still far below the OECD average of 485 (Wijaya et al., 2024). These findings suggest that many students still experience difficulties in understanding scientific concepts comprehensively and in applying them to solve contextual problems. Moreover, the proportion of Indonesian students who reach the minimum competency level remains relatively low compared to students in many other participating countries, reflecting ongoing challenges in improving the quality of the science education (Ayu et al., 2025; Setiaji et al., 2023). One contributing factor is that learning methods in schools are often still oriented toward memorization and routine exercises rather than conceptual understanding (Khotimah & Fariyani, 2025). As a result, students are less accustomed to solving contextual problems that require analysis and complex reasoning, which limits the development of higher-order thinking skills.

Several studies indicate that students' problem-solving skills in science learning are still in the low to moderate category. Research conducted by (Sinaga et al., 2023) and (Jatisunda et al., 2024) indicates many students experience difficulties in comprehending problems thoroughly, particularly when they must identify essential information presented in a question. This difficulty often occurs at the initial stage of problem-solving, which requires careful interpretation and analysis of the given problem. In addition, many students struggle to design appropriate and effective strategies to reach a correct solution, indicating that their reasoning processes have not developed optimally (Salim et al., 2024). To address this issue, systematic efforts are needed to train students' problem-solving skills through structured learning activities that emphasize analytical and reflective thinking (Brundage et al., 2024b; Mufadillah et al., 2024). One widely recognized framework for developing these skills is the problem-solving model proposed by Polya, which consists of four interconnected stages: understanding the problem, planning a strategy, implementing the plan, and reviewing the results obtained (Annam et al., 2024).

Through these stages, students are encouraged not only to focus on obtaining the correct final answer but also to understand the reasoning process involved in solving a problem. Therefore, the application of Polya's problem-solving indicators can serve as a strong foundation for developing assessment instruments that measure students' thinking processes more comprehensively (Meyer & Friege, 2026; Nabilah et al., 2025). Apart from instructional strategies, the assessment component is also essential in enhancing students' problem-solving skills. However, assessments commonly implemented in schools still tend to emphasize final results rather than evaluating the cognitive processes students experience during problem solving (Tanjung et al., 2024). Ideally, effective assessment should measure students' thinking stages holistically, starting from understanding the problem, planning strategies, implementing solutions, and evaluating the results. This condition highlights a gap between the demands of 21st-century learning and current

assessment practices, emphasizing the importance of developing assessment instruments capable of measuring students' problem-solving skills, particularly in the physics learning (Idrus et al., 2023; Sanjiartha et al., 2024).

Physics is a branch of science that requires strong analytical, logical, and reasoning skills in order to understand various natural phenomena systematically. One topic that has great potential to develop these abilities is heat, which is closely related to many phenomena encountered in everyday life, such as changes in the state of matter and heat transfer processes (Trisnasari & Oksiana, 2024). However, several studies reveal that many students still experience difficulties in understanding the concepts of temperature and heat and in solving problems related to these topics (Nurul, 2022). This condition indicates the need for evaluation instruments that are capable of measuring students' abilities more deeply and comprehensively. Recent studies emphasize the importance of developing assessment instruments that can measure higher-order thinking skills and students' ability to solve problems in the physics learning (Haris et al., 2025; Kusumaningtyas et al., 2024).

Systematic reviews in physics education indicate that improving students' problem-solving skills continues to be a major focus of recent research, as these skills are essential for understanding complex scientific phenomena and applying scientific concepts effectively (Brundage et al., 2024b; Nisa et al., 2024). Several studies in Indonesia have also developed HOTS-based physics assessment instruments and found that well-designed instruments can help identify students' conceptual difficulties and thinking processes more accurately (Brown & Singh, 2021; Simamora et al., 2024). However, many existing instruments are still designed for general physics topics and have not specifically focused on measuring problem-solving skills in particular materials, such as temperature and heat (Sukma et al., 2025). This condition indicates the need to develop more specific and systematically validated assessment instruments that can accurately measure students' problem-solving abilities in heat material. Developing high-quality testing instruments therefore becomes a strategic effort to support the improvement of physics education. Such instruments must be carefully designed and empirically tested to ensure that they meet the criteria of validity and reliability in a scientifically accountable manner (Sinurat et al., 2026).

Each test item must be designed with an appropriate level of difficulty and strong discriminating power so that it can effectively differentiate students' levels of ability (Brundage & Singh, 2023; Sherliyanti & Jauhariyah, 2024). Through well-constructed items, the developed instrument is expected to provide a more accurate and comprehensive description of students' problem-solving abilities. Therefore, this study aims to develop and analyze the quality of a knowledge test instrument used to measure students' problem-solving skills in heat material. The quality of the instrument is evaluated through several important aspects, including reliability, validity, item difficulty level, and item discrimination power. The instrument is also designed based on George Pólya's problem-solving indicators, which include four main stages: understanding the problem, devising a plan, carrying out the plan, and evaluating the results. Through the development of this instrument, it is expected that teachers will obtain a more effective assessment tool that can measure students' thinking processes and support the development of assessment systems aligned with the demands of 21st-century learning.

METHODS

The development procedure followed the ADDIE model, consisting of Analysis, Design, Development, Implementation, and Evaluation stages. This research is a type of research and development study. The test instrument developed consists of 10 essay questions compiled based on Polya's problem-solving skill indicators, namely understanding the problem, planning a solution strategy, implementing the solution strategy, and reviewing

the results. The analysis of the test instrument's quality is divided into two stages, namely theoretical and empirical analysis of the instrument's quality.

The theoretical quality of the instrument was analyzed using the theoretical validity of three expert validators, while the empirical analysis of the instrument quality was conducted by testing the validity, reliability, difficulty level, and discrimination based on the results of a limited trial by 33 first-year students enrolled in the Physics Education Study Program in 2025. Although the sample size was relatively small, it was appropriate for a limited trial aimed at evaluating the instrument's quality before broader implementation. Therefore, the focus was on establishing validity and reliability rather than generalizing the findings.

The theoretical analysis of the instrument's quality using theoretical validity was conducted by two expert lecturers from the Physics Department of Surabaya State University and one physics teacher using a 1–4 Likert scale covering the aspects of content, construct, and language appropriateness. From the validation results, the mode of each expert validator's score will be taken for each aspect. Then, the percentage of agreement between the three validators will be calculated using PoA (Percentage of Agreement) with the formula:

$$PoA = 100\% \times \left(1 - \frac{A-B}{A+B}\right) \quad \ldots (1)$$

Description:
A = Frequency of aspects rated by observers as high.
B = Frequency of aspects rated by observers as low.
According to Borich (1994), the acceptance criteria for the Percentage of Agreement (PoA) state that an instrument is considered valid when the obtained score reaches at least 75%.

Furthermore, empirical analysis of instrument quality is conducted based on the validity of the items calculated using the product-moment correlation as follow:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\{N\Sigma X^2 - (\Sigma X)^2\}\{N\Sigma Y^2 - (\Sigma Y)^2\}}} \quad \ldots (2)$$

Description:
r_{xy}= Product-moment correlation coefficient for each item
N = Number of test respondents
X = Score obtained by the subject on each item
Y = Total score obtained by the subject from all items
The criteria for testing the empirical validity of items were determined by comparing the r_{count} with r_{table} values. An item is declared valid if the r_{count} is greater than r_{table} (Arikunto, 2013).

Other than validity testing, reliability testing is carried out to assess the consistency of the instrument. The reliability test for essay tests can be calculated using the following Cronbach's alpha reliability formula (Arikunto, 2013):

$$r_{count} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma st^2}{st^2}\right) \quad \ldots (3)$$

Description:
r_{count} = Reliability Value
n = Number of items
Σst^2 = Total variance of each item score
st^2 = Total variance
From the formula above, the instrument will meet the reliability criteria if the calculated value is greater than 0.6, and conversely, the instrument does not meet the reliability criteria if the calculated value is less than 0.6.

In addition, a difficulty level test can be conducted to indicate how difficult or easy an item is. An item is considered good if it has a moderate level of difficulty, meaning that it

is neither too easy nor too difficult (Arikunto, 2013). This level of difficulty can be calculated using the following equation:

$$DL = \frac{Mean}{Maximum\ score} \quad \ldots (4)$$

Description:
DL = Difficulty level
Mean = Average score of students
Maximum score = Maximum score on the scoring rubric
The difficulty level of each item is determined using a difficulty index that ranges from 0.00 to 1.00. An item is considered acceptable if it falls within the moderate category, indicated by a difficulty index value between 0,31 and 0,70.

The final empirical analysis involves the discrimination power test, which refers to an item's ability to differentiate between test takers with high and low levels of ability (Arikunto, 2013). The discrimination power of an item can be calculated using the formula:

$$DP = \frac{\overline{X_A} - \overline{X_B}}{Maximum\ score} \quad \ldots (5)$$

Description:
DP = Discrimination power
$\overline{X_A}$ = Average score of the top group for each item
$\overline{X_B}$ = Average score of the bottom group for each item
Maximum score = Ideal maximum score

The discriminating power of an item is categorized based on a discrimination index ranging from 0.00 to 1.00. An item is considered to have adequate discriminating power when the discrimination index value is greater than or equal to 0.20.

RESULTS AND DISCUSSION

The development of the test instruments began with an analysis of needs and the initial conditions of the learning process. This stage involved examining the curriculum and basic competencies to determine the content to be used in the study. Based on the results of this analysis, the next step was to design the test instruments in alignment with the established indicators. The test instruments were developed based on problem-solving skill indicators referring to the established Polya model. In the design process, this test instrument is provided with a question outline, a question specification table, and an assessment rubric to ensure consistency between the indicators and the questions. Additionally, the test instrument is subjected to expert review and a small-scale trial to gather empirical evidence regarding its validity, reliability, level of difficulty, and discrimination index.

The classification of questions based on problem-solving skill indicators in accordance with Polya's model is presented in Table 1. The development process of this tool refers to four main stages, which are understanding the problem, devising the plan, carrying out the plan, and looking back. Each indicator is divided into several questions so that all aspects of problem-solving skills can be measured properly and systematically.

Table 1 Classification of question items based on polya's problem-solving skill indicators

No.	Polya model problem-solving skill indicators	Question items number
1.	Understanding the problem	1,4a, 6a, 8a, 10a
2.	Devising the plan	2a, 3a, 4b, 5a, 6b, 7a, 8b, 9a, 10b
3.	Carrying out the plan	2b, 4c, 6c, 8c, 10c
4.	Looking back	3b, 5b, 7b, 9b

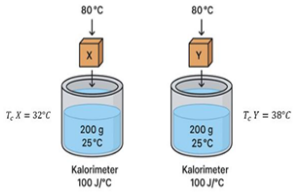
Based on the distribution in Table 1, the indicator of understanding the problem is represented by items 1, 4a, 6a, 8a, and 10a, which aim to measure students' ability to identify known and unknown information. The indicator of planning the solution is

represented by items 2a, 3a, 4b, 5a, 6b, 7a, 8b, 9a, and 10b, which require students to determine relevant strategies or equations before solving the problem. The indicator of implementing a solution plan includes questions 2b, 4c, 6c, 8c, and 10c, which focus on students' ability to apply solution steps and perform calculations accurately. Meanwhile, the indicator of checking the results is represented by items b, 5b, 7b, and 9b, which measure students' ability to review the process and evaluate the results obtained. The following are examples of items for each indicator of Polya's problem-solving skills as shown in Table 2.

In understanding the problem indicator, students are asked to identify known and unknown quantities and explain the process of heat transfer in the event of mixing hot tea and ice cubes. The devising the plan indicator is demonstrated through a question about heat conduction in an iron rod used to melt ice, where students are asked to develop steps to determine how long it takes to melt all the ice by determining the equation for the rate of heat transfer by conduction and the heat of fusion required. Furthermore, the indicator of carrying out the plan requires students to perform calculations to determine the equilibrium temperature based on calorimetry experimental data and a systematically written plan. Meanwhile, the indicator of reviewing the process and results obtained (looking back) is demonstrated through a question about choosing between ice cubes and cooling gel packs as cooling agents, in which students are asked to evaluate the strategies they have developed and conclude the effectiveness of their choices.

Table 2 Example of question items based on polya's problem-solving skill indicators

Polya model problem-solving skill indicators	**Example of question items**
Understanding the problem	 Arham pours 250 mL of hot tea (90°C) into a ceramic cup with a mass of 300 g at room temperature (25°C). In order to drink the tea immediately, Arham plans to add several ice cubes, where each cube has a mass of 25 g and an initial temperature of −5°C. However, Arham does not know how many ice cubes must be added so that the temperature of the tea decreases and reaches 40°C. (Given: specific heat capacity of tea = water = 4,200 J/kg°C, specific heat capacity of ceramic = 840 J/kg°C, specific heat capacity of ice = 2,100 J/kg°C, latent heat of fusion of ice = 336,000 J/kg, and density of water = 1 g/mL). 1. Based on the problem above, **identify the known physical quantities along with their units and determine what quantity is being asked in the problem**. Then, explain sequentially the heat transfer processes that occur between the hot tea, the ceramic cup, and the ice cubes until the entire system reaches a thermal equilibrium temperature of 40°C.
Devising the plan	6. One day, Andi wants to melt 1 kg of ice using an iron rod. He immerses one end of the iron rod in boiling water at 100°C, while the other end is in contact with ice at 0°C. The iron rod has a length of 50 cm and a diameter of 1 cm. It is assumed that all the heat flowing through the rod is used to melt the ice. From this event, Andi wants to know how long it takes for the 1 kg of ice to melt completely. (Given: thermal conductivity coefficient of iron, latent heat of fusion of ice = 336,000 J/kg). b. **Write the solution steps by formulating the equations** used to determine the required quantity based on the heat conduction rate equation through the iron rod, and determine the conduction rate and the heat required to melt 1 kg of ice

Polya model problem-solving skill indicators	Example of question items
Carrying out the plan	4. Observe the following illustration!  In a physics laboratory, a student conducts several experiments using a calorimeter with a heat capacity of 100 J/°C containing 200 g of water at 25°C. In the first experiment, 100 g of metal X at 80°C with a specific heat capacity of 1,370.8 J/kg°C is placed into the calorimeter, resulting in a mixture temperature of 32°C. In the second experiment, with the same initial conditions of the calorimeter and water, 100 g of metal Y at 80°C with a specific heat capacity of 2,909.5 J/kg°C is placed into the calorimeter and the final equilibrium temperature obtained is 38°C. Based on these experiments, the student wants to determine the final equilibrium temperature if metal X and metal Y, each initially at 80°C, are first attached together into one unit and then inserted simultaneously into the calorimeter under the same initial conditions as the previous experiments. c. **Use the solution steps written in point (b)** to perform the calculations systematically until the required value is obtained!
Looking back	9. A group of students will conduct a field trip to a mountainous area lasting an entire day, from morning until afternoon. During the activity, they plan to bring several cold beverages for the trip. A teacher provides two cooling options that can be used inside the drink box: ice cubes or cooling gel packs. The students must choose one of the cooling materials and arrange it properly inside the drink box so that the beverages remain cold and stable throughout the journey. b. **After deciding the solution strategy in point (a), make a conclusion** about whether the decision is appropriate and whether the chosen coolant is more effective than the other option.

The developed test instruments were then evaluated through theoretical validation by two lecturers and one physics teacher, focusing on aspects of content, construction, and language aspects using a validation sheet. The results of this validation for the problem-solving skills test instrument are shown in Table 3. This process is intended to examine the consistency of the items with the measured indicators, the accuracy of the questions, and the clarity of the language employed. Each aspect is judged based on the mode score, and the Percentage of Agreement (PoA) is used to determine the level of validity of the instrument.

Table 3 The results of theoretical validity of the test instrument

Aspect	Aspect number	Modus	Criteria
Construction	1	4	Very Good
PoA	85.71%		Valid
Content	1	3	Good
	2	4	Very Good
	3	3	Good
PoA	85.71%		Valid
Language	1	4	Very Good
	2	4	Very Good
	3	4	Very Good
	4	3	Good
PoA	92.85%		Valid
PoA Average	88.09%		Valid

Based on the validation results in Table 3, the results of theoretical validation showed that the developed instrument achieved an average Percentage of Agreement (PoA) of 88.09%, which falls into the valid category. In terms of construction, the PoA was 85.71% with a valid category and a rating mode of 4 (very good). The content aspect also received a PoA of 85.71% with a valid category, where most aspects had a rating mode of 3 (good). Meanwhile, the language aspect obtained a PoA of 92.85% with a valid category and a mode of 4 (very good), indicating that the use of language in the instrument was clear and in accordance with the rules. This indicates that the instrument has met the essential criteria in terms of content, construction, and language appropriateness. A high PoA value reflects a strong level of agreement among experts, suggesting that the items correspond to the intended problem-solving indicators based on Polya's framework.

This finding implies that the developed instrument has strong content validity, meaning that each item adequately represents the construct of problem-solving skills. According to Borich et al. (1994), a PoA value above 75% indicates acceptable agreement among validators, which supports the validity of the instrument. Furthermore, this result is consistent with previous studies on physics assessment instrument development, which emphasize the importance of expert validation to ensure alignment between indicators and test items (Aristiawan & Istiyono, 2020; Sherliyanti & Jauhariyah, 2024). From a practical perspective, the strong theoretical validity indicates that the instrument can serve as an appropriate tool for measuring students' problem-solving processes systematically. The alignment with Polya's stages ensures that the instrument not only evaluates final answers but also captures students' reasoning processes. From the theoretical validation results, the validators also provided input and recommendations for improving the developed instrument, which can be seen in Table 4.

Table 4 Suggestions and improvements provided by the validators

Suggestions and improvements	Question items number
Each question in the test instrument should be adjusted in more detail to the problem-solving skill indicators to be measured	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Overall, the recommendation based on Table 4 is that each question in the test should be adjusted in more detail to the problem-solving skill indicators to be measured. This recommendation applies to all questions, from number 1 to 10. This illustrates that even though the instrument already meets the validity criteria, improvements are still needed in terms of wording and content suitability so that each question accurately represents the indicators to be achieved. The follow-up to the validator's recommendation was carried out by revising the questions to make them more operational, focused, and in line with the stages in Polya's model. With these improvements, it is hoped that the instrument will have better alignment between the indicators, measurement objectives, and question characteristics. Furthermore, the test instrument that has been declared theoretically valid was tested on a limited sample of 33 first-year students enrolled in the Physics Education Study Program in 2025, as documented in Figure 1, to analyze its empirical validity. Empirical validity analysis of the test was carried out using Ms. Excel by comparing the r_{count} and r_{table}. values.

Figure 1 Test instrument testing documentation

Table 5 presents the results of the empirical validity analysis of the test items. This analysis aims to assess the degree to which each test item is able to measure the abilities that should be measured, based on the test results data

Table 5 The results of empirical validity

Question items	r_{count} Value	r_{table}	Criteria
1	0.505	0.344	Valid
2	0.621		Valid
3	0.572		Valid
4	0.642		Valid
5	0.754		Valid
6	0.794		Valid
7	0.690		Valid
8	0.741		Valid
9	0.572		Valid
10	0.782		Valid

Based on Table 5, it is known that the r_{table} value is 0.344. The empirical validity analysis revealed that all items (100%) were valid, as each item's correlation coefficient exceeded the critical value. This result indicates that all items have a significant relationship with the total score, suggesting that each item contributes meaningfully to measuring the same construct, namely problem-solving skills. This finding demonstrates that the developed instrument has strong construct validity, as each item consistently measures aspects of problem-solving skills. According to Arikunto (2013), items with correlation coefficients higher than the critical value can be considered valid because they effectively represent the measured variable. The consistency across all items suggests that the instrument has a well-structured design and coherent internal relationships among items.

Moreover, this result is in line with previous research showing that well-designed essay-based instruments are effective in measuring higher-order thinking skills, particularly problem-solving abilities in physics learning (Docktor & Mestre, 2014; Ekawati et al., 2024). The use of Polya's framework further strengthens the construct validity by ensuring that each item targets specific stages of the problem-solving process. Next, the results of the reliability analysis of the problem-solving skills test instrument were calculated using Ms. Excel, and the results are shown in Table 6. The reliability test was conducted to determine the extent of internal consistency between items in measuring the same construct, namely problem-solving skills based on Polya's model. Reliability was calculated using Cronbach's alpha coefficient because the instrument was descriptive and consisted of several interrelated items. Reliability is very important to ensure that the instrument provides stable and consistent results when used in measurement (Fadzil et al., 2022).

Table 6 The results of question items' reliability

r_{count} Value	Minimum value	Criteria
0.83	0.60	Reliable

Based on Table 6, the calculated r_{count} value was 0.83, which is greater than the minimum limit value, indicating a high level of reliability. This result suggests that the instrument has strong internal consistency, meaning that the items consistently measure the same construct. A reliability coefficient above 0.70 is generally considered acceptable, while values above 0.80 indicate high reliability (Arikunto, 2013; Shofiah et al., 2023). Therefore, the obtained value confirms that the instrument produces stable and consistent results when used in similar contexts. This consistency is particularly important in educational assessment, as reliable instruments ensure that the measurement results are dependable and not influenced by random errors.

The high reliability of the instrument also supports its use in both research and classroom assessment. It indicates that the instrument can provide accurate and consistent information about students' problem-solving skills, which is essential for evaluating learning outcomes and designing appropriate instructional interventions. In addition, an analysis of the level of difficulty of each item was also conducted to determine the proportion of the ease level of the questions that had been designed. The results of the difficulty level data processing can be seen in Table 7. This analysis aims to group the items into easy, medium, and difficult categories based on the results of the students' work. The difficulty level is one of the main indicators in evaluating the quality of the instrument because it can show the distribution of the abilities being measured (Novika et al., 2025).

Table 7 The results of difficulty level

Criteria	Question-number items	Total
Easy	1, 3, 5, 6, 7, 10	6
Medium	2, 4, 8, 9	4
Difficult	-	-

Based on Table 7, of the 10 questions analyzed, there were 6 questions (60%) that fell into the easy category, namely numbers 1, 3, 5, 6, 7, and 10. Furthermore, 4 items (40%) were classified as medium difficulty, namely numbers 2, 4, 8, and 9, while there were no items classified as difficult. These findings show that the majority of questions were easy, with the rest being medium difficulty. This distribution shows that the instrument measures more basic to intermediate understanding of students and does not contain many questions with a high level of difficulty. While the distribution of item difficulty is acceptable for initial validation purposes, the absence of difficult questions may make it harder to identify students with higher problem-solving skills. Adding more difficult items in future versions could improve the instrument's ability to differentiate among students of varying ability levels.

This finding suggests that the instrument may not fully capture higher-order problem-solving abilities, as more challenging items are typically required to assess complex reasoning and analytical skills. According to assessment theory, an ideal test should contain a balanced distribution of difficulty levels to differentiate students across a wide range of abilities (Brookhart, 2024). The dominance of easy items may indicate that students are more familiar with routine problem-solving tasks, but it also highlights the need to include more complex and non-routine problems that require deeper conceptual understanding. This is particularly important in physics learning, where problem-solving skills involve not only applying formulas but also analyzing situations and making decisions based on scientific reasoning (Docktor & Mestre, 2014). Therefore, future development of the instrument should consider adding more high-difficulty items to better measure higher-order thinking skills and offer a more thorough assessment of students' problem-solving abilities.

Then, an analysis of the discriminating power of the questions was conducted to determine the extent to which each question could distinguish between students with high and low levels of ability. Table 8 presents the results of the item discrimination analysis. This analysis aims to evaluate the extent to which the questions are able to identify differences in the mastery of problem-solving skills (Subhaktiyasa, 2024). Discrimination is an important indicator in determining the quality of items in an assessment instrument.

Table 8 The results of item discrimination power

Criteria	Question items Number	Total
Less	-	-
Enough	1, 2, 3, 4, 5, 6, 7, 8, 10	9
Good	9	1

According to Table 8, of the 10 questions analyzed, none of the questions had discriminating power in the low category. A total of 9 items (90%) were in the adequate category, including numbers 1, 2, 3, 4, 5, 6, 7, 8, and 10, while one item (10%), which was number 9, was in the good category. Items with good discrimination ability show an optimal ability to distinguish between hhigh-and low-ability students. On the other hand, items with discriminating power in the adequate category are still able to distinguish between students' abilities, although their accuracy is not yet fully maximized. However, overall, these results indicate that the developed instrument has adequate discriminating power and can be used to identify variations in students' problem-solving skills.

Items with good discrimination power are able to clearly differentiate between high-ability and low-ability students, while items with sufficient discrimination still function adequately but may require refinement. According to classical test theory, discrimination indices above 0.20 are considered acceptable, but higher values are preferred for more accurate differentiation (Arikunto, 2013). The predominance of items in the sufficient category suggests that the instrument is generally effective but still has room for improvement. Enhancing the quality of items, particularly by increasing their complexity and contextual relevance, may improve their ability to discriminate more sharply between different levels of student performance.

This finding aligns with previous studies indicating that the quality of test items significantly influences their ability to measure differences in students' abilities (Subhaktiyasa, 2024). Therefore, further refinement of the instrument is recommended to improve its discriminating power and overall effectiveness. Overall, the findings of this study indicate that the developed instrument meets the criteria of validity and reliability, making it suitable for assessing students' problem-solving skills on heat concepts. The integration of Polya's problem-solving framework provides a structured approach to assessing students' cognitive processes, which is essential for evaluating higher-order thinking skills.

However, the analysis also reveals several areas for improvement, particularly in terms of difficulty level and discrimination power. The dominance of easy items and the limited number of highly discriminative items suggest that further refinement is needed to enhance the instrument's ability to measure complex problem-solving skills. Nevertheless, the satisfactory validity and reliability results indicate that the instrument meets the basic standards for educational assessment. Although improvements in item difficulty and discrimination are still needed, the instrument remains useful for measuring students' problem-solving skills and can be further refined to assess a wider range of abilities. From an educational perspective, the developed instrument has significant implications for physics learning. It can be used as an alternative assessment tool that not only evaluates students' final answers but also captures their reasoning processes.

This is aligned with the goals of 21st-century education, which emphasize the development of critical thinking and problem-solving skills. Furthermore, the instrument can support teachers in identifying students' strengths and weaknesses in problem-solving, enabling more targeted instructional strategies. Therefore, the continued development and refinement of such instruments are essential for improving the quality of physics education and fostering students' higher-order thinking skills. The relatively small sample size used in this study may limit the generalizability of the findings to broader student populations. Although the sample was considered adequate for an initial instrument validation trial, future studies should involve larger and more diverse samples from different institutions to provide stronger evidence of the instrument's validity, reliability, and applicability across various educational contexts.

CONCLUSION

The developed knowledge test instrument used to measure problem-solving skills on the topic of heat shows strong theoretical validity. The theoretical validity results indicate that the construction aspect obtained a score of 85.71% in the valid category, the content aspect obtained 85.7% in the valid category, and the language aspect obtained 92.85% in the valid category. These results produced an average percentage of agreement of 88.09%, which is categorized as valid. In the empirical validity test, all test items, representing 100% of the items, were declared valid because the r_{count} values were higher than the r_{table} values. The reliability coefficient obtained was 0.83, indicating that the instrument is reliable. Based on the difficulty level analysis, 60% of the items were categorized as easy, 40% were categorized as moderate, and none of the items were categorized as difficult.

Furthermore, the discrimination power analysis showed that 90% of the items were classified as acceptable, while the remaining 10% were categorized as good, with no items categorized as poor. Therefore, it can be inferred that the developed test instrument is valid and reliable and possesses appropriate item characteristics to measure students' problem-solving skills, making it feasible for use in learning activities and further research, as well as a reference for designing more systematic physics assessments aligned with curriculum demands. Thereby contributing to preparing students to face future learning challenges and scientific problems. However, this study is limited by a small and context-specific sample, a question difficulty distribution dominated by easy and moderate levels without difficult items, and mostly adequate discriminative power, suggesting the need for future research to involve more diverse samples, include a wider range of difficulty levels, and improve item discrimination to enhance the instrument's overall effectiveness.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

ATY: Data Curation, Project Administration, Methodology, Formal Analysis, Resources, and Writing - Original Draft. MNRJ: Conceptualization, Methodology, and Validation.

AI DISCLOSURE STATEMENT

The authors used ChatGPT during the preparation of this manuscript solely to support idea generation and initial structuring of the content. After using this tool, the authors thoroughly reviewed, revised, and refined the entire manuscript to ensure its accuracy, coherence, and academic integrity. The authors take full responsibility for the final content. All aspects of the research process, including research design, data collection, data analysis, and drawing conclusions, were carried out independently without reliance on artificial intelligence (AI).

REFERENCES

Adeoye, M. A., & Jimoh, H. A. (2023). Problem-solving skills among 21st-century learners toward creativity and innovation ideas. *Thinking Skills and Creativity Journal*, *6*(1), 52–58. https://doi.org/10.23887/tscj.v6i1.62708

Annam, S., Syuzita, A., Pratiwi, R., & Sarkingobir, Y. (2024). Mapping research trends on 21st-century problem-solving skills in science learning: A literature review from 2014 to 2023. *Research in Education, Technology, and Multiculture*, *3*(1), 52–60. https://doi.org/10.61436/rietm/v3i1.pp52-60

Arikunto, S. (2013). *Prosedur penelitian*. Jakarta: Rineka Cipta.

Aristiawan, A., & Istiyono, E. (2020). Developing instrument of essay test to measure the problem-solving skill in physics. *Jurnal Pendidikan Fisika Indonesia*, *16*(2), 72–82. https://doi.org/10.15294/jpfi.v16i2.24249

Ayu, G. N., Putri, C. A., Riyanto, A. R., & Koto, I. (2025). The scientific literacy competence of students in indonesia and mexico based on pisa 2022: An international comparative study. *TOFEDU: The Future of Education Journal*, *4*(5), 1033–1038. https://doi.org/10.61445/tofedu.v4i5.525

Bhakti, Y. B., Arthur, R., & Supriyati, Y. (2025). Development of physics hots assessment instruments in high school: A comprehensive literature review. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *10*(2), 194. https://doi.org/10.26737/jipf.v10i2.5962

Borich, G. D., Godbout, R. C., Peck, R. F., Kash, M. M., & Poynor, L. H. (1994). An evaluation of the personalized model of teacher. *Research and Development Center for Teacher Education The University of Texas at Austin*.

Brookhart, S. M. (2024). Educational Assessment knowledge and skills for teachers revisited. *Education Sciences*, *14*(7). https://doi.org/10.3390/educsci14070751

Brown, B., & Singh, C. (2021). Development and validation of a conceptual survey instrument to evaluate students' understanding of thermodynamics. *Physical Review Physics Education Research*, *17*(1). https://doi.org/10.1103/PhysRevPhysEducRes.17.010104

Brundage, M. J., Meltzer, D. E., & Singh, C. (2024a). Investigating introductory and advanced students' difficulties with change in internal energy, work, and heat transfer using a validated instrument. *Physical Review Physics Education Research*, *20*(1). https://doi.org/10.1103/PhysRevPhysEducRes.20.010115

Brundage, M. J., Meltzer, D. E., & Singh, C. (2024b). Investigating introductory and advanced students' difficulties with entropy and the second law of thermodynamics using a validated instrument. *Physical Review Physics Education Research*, *20*(2), 1–21. https://doi.org/10.1103/PhysRevPhysEducRes.20.020110

Brundage, M. J., & Singh, C. (2023). Development and validation of a conceptual multiple-choice survey instrument to assess student understanding of introductory thermodynamics. *Physical Review Physics Education Research*, *19*(2), 20112. https://doi.org/10.1103/PhysRevPhysEducRes.19.020112

Docktor, J. L., & Mestre, J. P. (2014). A synthesis of discipline-based education research in physics. *Physical Review Special Topics-Physics Education Research*, *10*(Special Topics), 1-148 (020119). https://doi.org/10.1103/PhysRevSTPER.10.020119

Ekawati, E. Y., Azzahra, S. A., & Fithriyah, H. (2024). Construction of critical reasoning skills assessment instruments as diagnostic assessments in physics learning with polytomus scoring. *Journal of Education Research and Evaluation*, *8*(2), 338–349. https://doi.org/10.23887/jere.v8i2.69259

Fadzil, H. M., Saat, R. M., & Rafi, A. (2022). Development of technology-enhanced three-tier diagnostic test to assess pre-university students' understanding of scientific concepts. *Alternative Assessments in Malaysian Higher Education: Voices from the Field*, 285–292. https://doi.org/10.1007/978-981-16-7228-6_29

Hadi, H., Juandi, D., & Rusdiana, D. (2023). Problem solving ability analysis: Systematic literature review. *Journal of Mathematics and Mathematics Education, 13(1), 33-43*. https://doi.org/10.20961/jmme.v13i1.73819

Haris, A., Hasyim, M., Ma'ruf, Halim, A. D., Mahir, & Ramadhan, I. (2025). The effectiveness of dynamic problem-solving strategy and innovative assessment tools to improve students' higher-order thinking skills. *Jurnal Pendidikan IPA Indonesia*, *14*(1). https://doi.org/10.15294/jpii.v14i1.21433

Hidayati, F., & Julianto, J. (2025). Integrasi pendekatan etnosains dalam pembelajaran sains untuk meningkatkan keterampilan berpikir kritis siswa. *DIDAKTIKA : Jurnal Pemikiran Pendidikan*, *31*(1), 101. https://doi.org/10.30587/didaktika.v31i1.9578

Idrus, T., Pomalato, S. W., Ismail, S., & Mohidin, A. D. (2023). Pengembangan instrumen kemampuan pemecahan masalah matematika peserta didik di sma. *Jambura Journal of Mathematics Education, 4*(1), 63–74. https://doi.org/10.34312/jmathedu.v4i1.16404

Jatisunda, M. G., Suryadi, D., & Prabawanto, S. (2024). Enhancing mathematical problem-solving ability through project-based learning: A study of vocational high school student. *Jurnal Elemen*, *10*(3), 711–727. https://doi.org/10.29408/jel.v10i3.26244

Khotimah, K., & Fariyani, Q. (2025). Development of a physics assessment instrument to measure high school students multirepresentation skills in momentum and impulse. *Impulse: Journal of Research and Innovation in Physics Education*, *5*(1), 49–61. https://doi.org/10.14421/impulse.2025.51-04

Kusumaningtyas, D. A., Manyunu, M., Kurniasari, E., Awalin, A. N., Rahmaniati, R., & Febriyanti, A. (2024). Enhancing learning outcomes: A study on the development of higher order thinking skills based evaluation instruments for work and energy in high school physics. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, 14–31. https://doi.org/10.23917/ijolae.v6i1.23125

Latifah, S., Meisuri, M., & Kesuma, A. A. (2023). Development of physics worksheet based on higher order thinking skills (hots) integrated with bilingual class learning. *Indonesian Journal of Science and Mathematics Education*, *6*(1), 82–95.

Lubis, M. U., Siagian, F. A., Zega, Z., Nuhdin, N., & Nasution, A. F. (2023). Pengembangan kurikulum merdeka sebagai upaya peningkatan keterampilan abad 21 dalam pendidikan. *ANTHOR: Education and Learning Journal*, *2*(5), 691–695. https://doi.org/10.31004/anthor.v1i5.222

Meyer, A., & Friege, G. (2026). An atomized approach to assessing energy problem solving in physics using multidimensional item response theory. *Frontiers in Education*, *11*(February). https://doi.org/10.3389/feduc.2026.1759878

Mufadillah, M., Trimurtini, T., & Supriyanto, T. (2024). Profile of problem-solving skills question story based on the polya model at kampus mengajar partner schools. *Numerical: Jurnal Matematika Dan Pendidikan Matematika*, *8*(1), 187–204. https://doi.org/10.25217/numerical.v8i1.4379

Nabilah, N., Umma, D. K., Nasution, S. M., & Putri, M. D. (2025). Pengembangan instrumen tes essay berbasis kemampuan berpikir kritis materi suhu dan kalor pada pembelajaran ipa di sekolah menengah pertama. *Jurnal Inovasi Pendidikan Sains (JIPS)*, *6*(2), 74–81. https://doi.org/10.37729/jips.v6i2.5999

Nisa, K., Putri, S. A., & Kuswanto, H. (2024). A systematic literature review: Problem-solving skills in physics teaching. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, *10*(1), 173–182. https://doi.org/10.21009/1.10115

Novika, R., Istiyono, E., & Notiavina, A. A. (2025). Developing an assessment instrument to measure the minimum competency of high school physics using item response theory. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *29*(1), 98–113. https://doi.org/10.21831/pep.v29i1.70700

Nurul, D. (2022). Analisis kesulitan kemampuan pemecahan masalah pada pesera didik dalam pembelajaran fisika. *Jurnal Inovasi Dan Teknologi Pendidikan*, *1*(1), 20–30. https://doi.org/10.46306/jurinotep.v1i1.2

Osborne, J., & Allchin, D. (2025). Science literacy in the twenty-first century: Informed trust and the competent outsider. *International Journal of Science Education*, *47*(15–16), 2134–2155. https://doi.org/10.1080/09500693.2024.2331980

Salim, F., Purwanto, A., & Lestari, I. (2024). Improving students' science problem solving ability through the implementation of problem based learning models assisted by animation media. *International Journal of Elementary Education*, *8*(2), 269–278. https://doi.org/10.23887/ijee.v8i2.76925

Sanjiartha, I. G. D., Suwindia, I. G., & Winangun, I. M. A. (2024). Peran literasi sains dalam membentuk generasi berfikir kritis dan inovatif: Kajian literature review. *Education and Social Sciences Review*, *5*(2), 120–128. https://doi.org/10.29210/07essr499900

Setiaji, B., Santoso, P. H., Aziz, K. N., Retnawati, H., & Khairudin, M. (2023). Using

multilevel modelling to evaluate science literacy and technology course of the indonesian non-science students. *Jurnal Pendidikan IPA Indonesia*, *12*(1), 96–111. https://doi.org/10.15294/jpii.v12i1.41457

Sherliyanti, Y., & Jauhariyah, M. N. R. (2024). Validitas instrumen penilaian berbasis kelas untuk mengukur kompetensi literasi sains peserta didik pada materi pemanasan global. *IPF: Inovasi Pendidikan Fisika*, *13*(3), 168–176. https://doi.org/10.26740/ipf.v13n3.p168-176

Shofiah, S., Bachtiar, E., & Kd, D. P. (2023). *Dasar-dasar evaluasi pembelajaran*. Deli Serdang: PT. Mifandi Mandiri Digital.

Simamora, N. N., Istiyono, E., & Wilujeng, I. (2024). Development of an assessment instrument to measure high school students' collaboration skills in physics learning. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, *10*(1). https://doi.org/10.21009/1.10104

Sinaga, B., Sitorus, J., & Situmeang, T. (2023). The influence of students' problem-solving understanding and results of students' mathematics learning. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1088556

Sinurat, J. N., Safriana, Ayunda, D. S., Ginting, F. W., & Andriani, R. (2026). Evaluating assessment instruments for measuring students' higher-order thinking skills in heat and temperature concepts. *Jurnal Phi Jurnal Pendidikan Fisika Dan Fisika Terapan*, *12*(1), 54–65. https://doi.org/10.22373/p-jpft.v12i1.32237

Subhaktiyasa, P. G. (2024). Evaluasi validitas dan reliabilitas instrumen penelitian kuantitatif: sebuah studi pustaka. *Journal of Education Research*, *5*(4), 5599–5609. https://doi.org/10.37985/jer.v5i4.1747

Sukma, M. D. A., Dhiarni, Y. I., Kusniati, R., Wiyatmo, Y., & Kuswanto, H. (2025). Systematic literature review: Analysis of problem-solving skill trends in physics education (2014-2024). *Jurnal Riset Dan Kajian Pendidikan Fisika*, *12*(1), 41–49. https://doi.org/10.12928/jrkpf.v12i1.1145

Tanjung, Y. I., Sanjaya, A. P., & Mihardi, S. (2024). Development of physics test instruments to measure problem solving skills on dynamic fluid materials. *Journal of Innovative Physics Teaching*, *2*(2), 89–101. https://doi.org/10.24036/jipt/vol2-iss2/68

Trisnasari, E., & Oksiana, S. (2024). Identification of students' conceptual understanding of temperature and heat material. *EduFisika: Jurnal Pendidikan Fisika*, *9*(3), 260–267. https://doi.org/10.59052/edufisika.v9i3.37601

Wijaya, T. T., Hidayat, W., Hermita, N., Alim, J. A., & Talib, C. A. (2024). Exploring contributing factors to pisa 2022 mathematics achievement: insights from indonesian teachers. *Infinity Journal*, *13*(1), 139–156. https://doi.org/10.22460/infinity.v13i1.p139-156

AUTHOR PROFILES

Adhelia Tasya Youlanda
Email: adheliatasya.22006@mhs.unesa.ac.id
Affiliation: Physics Education Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, Indonesia.

Dr. Mukhayyarotin Niswati Rodliyatul Jauhariyah, S.Pd., M.Pd.
Email: mukhayyarotinjauhariyah@unesa.ac.id
Affiliation: Physics Education Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, Indonesia.
SINTA ID: 74493; Scopus ID: 57128452400
https://scholar.google.com/citations?user=57kiu0gAAAAJ&hl=en