Berkala Ilmiah Pendidikan
Volume 6, Number 3, November 2026

scidac plus

Assessing Students' Problem-Solving Skills in Heat Learning: Preliminary Validation of a Five-Stage Physics Assessment Instrument

Ardian Fachreza*, Andhy Setiawan, Muslim
Universitas Pendidikan Indonesia, Bandung, Indonesia
ardianfachreza@upi.edu*

Abstract

Conventional physics assessments often emphasize final numerical answers and therefore provide limited information about where students' reasoning breaks down during problem solving. This study developed and preliminarily validated a process-oriented assessment instrument for heat learning. A quantitative preliminary instrument-validation design was used. Four contextual essay problems were decomposed into five stages focusing on the problem, describing the physics, planning the solution, executing the plan, and evaluating the answer yielding 20 scored sub-items using a 0–4 analytic rubric. Five experts evaluated content validity using I-CVI, S-CVI/Ave, and S-CVI/UA, followed by an empirical trial with 30 Grade 12 students. Empirical analyses included corrected item-total correlations, Cronbach's alpha, item difficulty, and item discrimination. The instrument obtained S-CVI/Ave = 0.960 and S-CVI/UA = 0.800, with I-CVI values ranging from 0.80 to 1.00. Corrected item-total correlations ranged from 0.605 to 0.909, Cronbach's alpha was 0.975, difficulty indices ranged from 0.583 to 0.750, and discrimination indices ranged from 0.406 to 0.750. These findings provide preliminary evidence that the instrument can support stage-specific diagnosis of students' problem-solving performance in heat learning. Larger-sample validation and inter-rater reliability studies are still required.

Keywords: Assessment Instrument, Heat Learning, Physics Assessment, Problem-Solving Skills, Validation

INTRODUCTION

Problem solving is a central goal of physics education because successful solutions require more than recalling formulas. Students must represent the situation, identify relevant physical principles, plan a route to the solution, carry out the mathematics, and judge whether the result is physically plausible. Research on physics problem solving has long shown that expert performance is characterized by structured representations and principled reasoning rather than immediate equation selection (Heller & Reif, 1984; Heller et al., 1992).

Assessment practices that emphasize only a final numerical answer can therefore hide important differences in students' reasoning. A correct result may follow from incomplete or poorly justified reasoning, while a conceptually sound solution may contain a minor computational error. Process-oriented rubrics address this limitation by scoring observable components of a written solution rather than reducing performance to a single correct/incorrect judgment (Docktor et al., 2016).

One influential approach organizes physics problem solving into a sequence that requires students to focus on the problem, describe the physics, plan a solution, execute the plan, and evaluate the answer. Related work has shown that explicit problem-solving structures can serve both as instructional scaffolds and as assessment frameworks for examining how students approach complex problems (Heller et al., 1992; Burkholder et al., 2020). This staged perspective is particularly useful when the purpose of assessment is diagnostic rather than merely summative.

Heat and thermal physics provide an appropriate context for such diagnosis because students frequently have difficulty distinguishing concepts such as heat, temperature, and internal energy and may retain alternative conceptions even after instruction. Diagnostic studies in thermal physics have documented persistent conceptual difficulties and have emphasized the value of assessments that reveal the reasoning behind students' responses (Leinonen et al., 2013; Gurcay & Gulbas, 2015).

The instrument examined in this study applies the five-stage problem-solving structure to four heat contexts: basic heat transfer during heating, phase change, thermal expansion, and thermal equilibrium using the principle of mixtures (Black's principle). Each problem is divided into five scored stages, producing 20 sub-items. The analytic structure is intended to distinguish difficulties in identifying information and principles from weaknesses in planning, calculation, or evaluation of the final answer.

Before such stage-specific scores can be interpreted, evidence is needed that the instrument content is appropriate and that the scored components function consistently. Content validity can be examined through expert judgment using item and scale-level content validity indices, whereas empirical item analysis can provide preliminary information about sub-item relationships, score consistency, difficulty, and discrimination (Polit & Beck, 2006; Polit et al., 2007; Crocker & Algina, 1986). Cronbach's alpha can describe internal consistency, but it should not be treated as proof that all stages measure a single dimension (Taber, 2018).

Accordingly, this study provides preliminary validation evidence for a five-stage problem-solving assessment in heat learning. The study addresses three questions: (1) What evidence of content validity is obtained from expert judgments? (2) What initial evidence of score consistency and corrected item-total relationships is obtained from the student trial? and (3) What are the difficulty and discrimination characteristics of the 20 scored sub-items? The results are intended as an initial basis for further validation rather than as confirmation of the instrument's full dimensional structure.

METHOD

This study used a quantitative preliminary instrument-validation design to obtain initial evidence about content relevance, internal consistency, and item functioning. The validation was conducted in two stages: expert review of the instrument content followed by an empirical trial with students. This sequence is consistent with instrument-development practice in which evidence based on test content is established before score-based item analysis is interpreted (Lynn, 1986; Polit et al., 2007; Crocker & Algina, 1986). Because the present sample was small, the empirical analyses were treated as exploratory and were not used to claim confirmation of the instrument's dimensional structure.

The research was conducted at a senior high school in Bandung, Indonesia. Content validation was carried out by five validators with expertise in physics education and assessment. The empirical trial involved 30 Grade 12 students who had previously completed instruction on heat and thermodynamics. Prior exposure to the relevant content domain served as the inclusion criterion, ensuring that participants possessed sufficient conceptual knowledge to engage meaningfully with the assessment tasks. The instrument was administered under supervised classroom conditions following the completion of content validation and subsequent revisions.

The assessment instrument was a contextual essay test designed to make students' problem-solving processes visible the context of heat learning. It comprised four main problems: heating water, melting ice, thermal expansion of a railway track, and mixing hot and cold water using Black's principle. Each problem followed five stages, focusing the problem, describing the physics, planning the solution, executing the plan, and evaluating the answer, adapted from established structured approaches to physics problem solving (Heller et al., 1992; Burkholder et al., 2020). The repeated structure yielded 20 scored sub-items, labelled 1a through 4e. Tables 1 and 2 summarize the problem contexts and assessed stages.

Table 1. Structure of the Problem-Solving Assessment Instrument

No.	Physics Context	Main Target
1	Basic concept of heat	Determine the heat absorbed during heating
2	Effect of heat on phase change	Determine the latent heat required for melting
3	Thermal expansion in everyday life	Determine the increase in railway-track length
4	Black's principle	Determine the final temperature of mixed water

Table 2. Five Stages of the Assessment

Stage	Assessment Focus
Focus the problem	Identify relevant known and unknown information
Describe the physics	Select and express the relevant physical principle or equation
Plan the solution	Construct a logical sequence of solution steps
Execute the plan	Substitute quantities and complete the calculation
Evaluate the answer	Check units, reasonableness, completeness, and contextual meaning

Each sub-item was scored with an analytic rubric ranging from 0 to 4. A score of 0 indicated no scorable response; scores of 1 and 2 represented substantially incomplete or incorrect performance; a score of 3 represented generally appropriate performance with a minor omission or error; and a score of 4 represented complete and contextually appropriate performance for the stage. Analytic scoring was used because physics problem-solving rubrics can distinguish meaningful components of written reasoning that are not visible from the final numerical answer alone (Docktor et al., 2016).

Content validity was evaluated by five validators with expertise in physics education and assessment. For each of the 20 sub-items, validators judged three components: alignment between the problem-solving indicator and the item, appropriateness of the answer key, and appropriateness of the scoring rubric. Ratings were recorded as appropriate or not appropriate. The Item Content Validity Index (I-CVI) was calculated as the proportion of validators endorsing a component as appropriate. Scale-level content validity was summarized using S-CVI/Ave, the average of the I-CVIs, and S-CVI/UA, the proportion of components receiving universal agreement. Following Polit et al. (2007), I-CVI values of 0.78 or higher were treated as adequate preliminary evidence. For scale-level interpretation, an S-CVI/Ave of approximately 0.90 and an S-CVI/UA of approximately 0.80 were used as reference benchmarks (Polit & Beck, 2006). Validator comments were also used to revise the instrument before the student trial.

After revision, the instrument was administered to 30 Grade 12 students. Corrected item-total correlations were calculated to examine the association between each sub-item and the total score formed from the remaining sub-items. The study used 0.361 as its prespecified reference value for $n = 30$ at the 5% level. Internal consistency across the 20 polytomously scored sub-items was estimated with Cronbach's alpha. Alpha was interpreted as an index of score consistency, not as evidence of unidimensionality (Tavakol & Dennick, 2011; Taber, 2018).

For each sub-item, the difficulty index was calculated as the mean student score divided by the maximum score of four. The prespecified categories were difficult ($P < 0.30$), moderate ($0.30 \leq P \leq 0.70$), and easy ($P > 0.70$). Item discrimination was estimated from the difference between the mean scores of the upper and lower 27% groups; with 30 participants, each extreme group contained eight students. The 27% extreme-group approach follows the classical item-analysis procedure proposed by Kelley (1939), while the overall interpretation of difficulty and discrimination was treated as preliminary because these statistics are sample dependent (Crocker & Algina, 1986). Decisions about retaining sub-items considered the item statistics together with content-validity evidence and validator feedback.

RESULTS AND DISCUSSION

Results

Content Validity

The content-validation stage produced quantitative evidence regarding the alignment between the assessment components and the intended problem-solving constructs. Five validators evaluated three aspects for each of the 20 sub-items: the correspondence between the problem-solving indicator and the assessment item, the appropriateness of the answer key, and the appropriateness of the scoring rubric. The results are summarized in Table 3.

Table 3. Content Validity Indices of the 20 Problem-Solving Sub-items

Sub-Item	I-CVI (Indicator–Item)	I-CVI (Answer Key)	I-CVI (Rubric)	Decision
1a	1.00	1.00	0.80	Meets criterion; minor rubric revision
1b	1.00	1.00	0.80	Meets criterion; minor rubric revision
1c	1.00	1.00	0.80	Meets criterion; minor rubric revision
1d	1.00	1.00	1.00	Meets criterion
1e	1.00	1.00	1.00	Meets criterion
2a	1.00	1.00	0.80	Meets criterion; minor rubric revision
2b	1.00	1.00	0.80	Meets criterion; minor rubric revision
2c	1.00	1.00	0.80	Meets criterion; minor rubric revision
2d	1.00	1.00	1.00	Meets criterion
2e	1.00	1.00	1.00	Meets criterion
3a	1.00	1.00	0.80	Meets criterion; minor rubric revision
3b	1.00	1.00	0.80	Meets criterion; minor rubric revision
3c	1.00	1.00	0.80	Meets criterion; minor rubric revision
3d	1.00	1.00	1.00	Meets criterion
3e	1.00	1.00	1.00	Meets criterion
4a	1.00	1.00	0.80	Meets criterion; minor rubric revision

Sub-Item	I-CVI (Indicator–Item)	I-CVI (Answer Key)	I-CVI (Rubric)	Decision
4b	1.00	1.00	0.80	Meets criterion; minor rubric revision
4c	1.00	1.00	0.80	Meets criterion; minor rubric revision
4d	1.00	1.00	1.00	Meets criterion
4e	1.00	1.00	1.00	Meets criterion

As shown in Table 3, all indicator-to-item correspondence components and all answer-key components obtained perfect agreement among the five validators, yielding I-CVI = 1.00 across all 20 sub-items. For the scoring rubric, a consistent pattern emerged: sub-items representing the first three stages (focusing the problem, describing the physics, and planning the solution) obtained I-CVI = 0.80, indicating that four of five validators rated these rubrics as appropriate. Sub-items representing the fourth and fifth stages (executing the plan and evaluating the answer) obtained I-CVI = 1.00, reflecting unanimous agreement.

At the scale level, S-CVI/Ave was 0.960 and S-CVI/UA was 0.800. The S-CVI/Ave exceeded the 0.90 reference level for the averaging approach, while the S-CVI/UA met the traditional 0.80 benchmark for universal agreement (Polit & Beck, 2006). The scoring-rubric components with I-CVI = 0.80 were revised in response to validator comments before the empirical administration.

Empirical Item Characteristics

Following content validation and revision, the instrument was administered to 30 Grade 12 students. Each student's responses were scored on the 0–4 analytic rubric for each of the 20 sub-items. Corrected item-total correlations were computed to examine the relationship between each sub-item and the remaining total score. With $n = 30$ and a significance level of 5% ($df = 28$), the critical reference value was 0.361. The complete empirical characteristics of all 20 sub-items are presented in Table 4.

Table 4. Empirical Characteristics of the 20 Problem-Solving Sub-items

Sub-Item	Item-Total r	P	Difficulty	D	Discrimination	Decision
1a	0.679	0.725	Easy	0.406	Good	Used
1b	0.605	0.600	Moderate	0.438	Good	Used
1c	0.757	0.708	Easy	0.500	Good	Used
1d	0.846	0.608	Moderate	0.750	Very good	Used
1e	0.896	0.692	Moderate	0.750	Very good	Used
2a	0.785	0.625	Moderate	0.562	Good	Used
2b	0.828	0.658	Moderate	0.562	Good	Used
2c	0.801	0.583	Moderate	0.562	Good	Used
2d	0.877	0.625	Moderate	0.750	Very good	Used
2e	0.909	0.683	Moderate	0.625	Good	Used
3a	0.832	0.700	Moderate	0.594	Good	Used
3b	0.839	0.650	Moderate	0.656	Good	Used

Sub-Item	Item-Total r	P	Difficulty	D	Discrimination	Decision
3c	0.822	0.717	Easy	0.562	Good	Used
3d	0.863	0.717	Easy	0.750	Very good	Used
3e	0.861	0.675	Moderate	0.750	Very good	Used
4a	0.770	0.733	Easy	0.500	Good	Used
4b	0.806	0.692	Moderate	0.562	Good	Used
4c	0.627	0.650	Moderate	0.469	Good	Used
4d	0.823	0.733	Easy	0.531	Good	Used
4e	0.840	0.750	Easy	0.750	Very good	Used

All 20 sub-items produced corrected item-total correlation coefficients exceeding the criterion value of 0.361. The coefficients ranged from 0.605 (sub-item 1b) to 0.909 (sub-item 2e). Even the lowest coefficient remained substantially above the predetermined threshold, indicating that all sub-items were positively associated with the overall assessment score.

Internal Consistency

Cronbach's alpha for the 20 sub-items was 0.975. This indicates very high internal consistency of the scores in this sample, although the coefficient alone does not establish that the five stages are unidimensional (Tavakol & Dennick, 2011; Taber, 2018).

Item Difficulty

The difficulty index for each sub-item was calculated by dividing the mean student score by the maximum possible score of four. The resulting indices ranged from 0.583 (sub-item 2c) to 0.750 (sub-item 4e). Based on the predetermined classification criteria, 13 sub-items were categorized as moderate and seven sub-items were categorized as easy. The seven easy sub-items were 1a, 1c, 3c, 3d, 4a, 4d, and 4e. No sub-item was classified as difficult. The distribution of difficulty indices is presented in Table 5.

Table 5. Distribution of Item Difficulty

Difficulty Category	Number of Sub-Item	Sub-Item
Difficult ($P < 0.30$)	0	—
Moderate ($0.30 \leq P \leq 0.70$)	13	1b, 1d, 1e, 2a, 2b, 2c, 2d, 2e, 3a, 3b, 3e, 4b, 4c
Easy ($P > 0.70$)	7	1a, 1c, 3c, 3d, 4a, 4d, 4e

Item Discrimination

Item discrimination was estimated by comparing the mean scores of the upper 27% and lower 27% student groups. With 30 participants, each extreme group comprised eight students. The discrimination indices ranged from 0.406 (sub-item 1a) to 0.750 (sub-items 1d, 1e, 2d, 3d, 3e, and 4e). Fourteen sub-items were categorized as good ($0.40 \leq D < 0.70$), and six sub-items were categorized as very good ($D \geq 0.70$). No sub-item exhibited low or negative discrimination. The distribution is summarized in Table 6.

Table 6. Distribution of Item Discrimination

Discrimination Category	Number of Sub-Item	Sub-Item
Very good ($D \geq 0.70$)	6	1d, 1e, 2d, 3d, 3e, 4e
Good ($0.40 \leq D < 0.70$)	14	1a, 1b, 1c, 2a, 2b, 2c, 2e, 3a, 3b, 3c, 4a, 4b, 4c, 4d
Low ($0.20 \leq D < 0.40$)	0	—
Negative ($D < 0.20$)	0	—

Discussion

Taken together, the content and empirical results support the instrument as a preliminary diagnostic tool, but they also identify areas that require further validation. The strongest evidence is the high level of expert agreement: all indicator–item and answer-key components received I-CVI = 1.00, while the rubric components for the first three stages received I-CVI = 0.80. The resulting S-CVI/Ave of 0.960 and S-CVI/UA of 0.800 meet the respective reference levels commonly used for the two scale-level CVI approaches (Polit & Beck, 2006; Polit et al., 2007).

The lower agreement on rubric descriptors for focusing, describing, and planning is substantively important. These stages require judgments about the quality of students' representations and reasoning, whereas execution is more directly observable in mathematical work. Similar physics-assessment research has shown that analytic rubrics can make problem-solving processes visible, but the usefulness of such rubrics depends on clearly specified scoring criteria (Docktor et al., 2016). The validator revisions made before the student trial were therefore necessary rather than merely editorial.

All corrected item-total correlations exceeded the study's prespecified criterion, suggesting that the 20 sub-items were positively associated with the total score in this sample. The very high Cronbach's alpha (0.975) likewise indicates strong internal consistency. However, alpha increases when items measure overlapping content or repeated task structures, so a coefficient of this magnitude should not be interpreted as proof that the five stages form one psychological dimension (Tavakol & Dennick, 2011; Taber, 2018). The repeated use of the same five stages across four problems may itself contribute to the high coefficient.

The difficulty results show that most sub-items were moderate and the remainder were easy, with no difficult sub-items. For a process-oriented diagnostic instrument, this distribution has an advantage: students are sufficiently able to attempt each stage, allowing their written reasoning to be scored. At the same time, the absence of difficult sub-items may reduce differentiation among stronger students. Future versions should therefore test additional problems with more demanding conceptual or multistep reasoning while retaining the same stage structure.

Discrimination was positive for every sub-item, ranging from 0.406 to 0.750. Six sub-items reached the study's "very good" category, and five of these represented the execution or evaluation stages. This pattern suggests that later stages may be especially sensitive to differences in overall problem-solving performance because successful execution and evaluation depend on information and decisions made earlier in the solution. Process-based frameworks in physics similarly emphasize the importance of linking representation, strategy, calculation, and checking rather than treating them as unrelated actions (Heller et al., 1992; Burkholder et al., 2020).

The present instrument also differs from common diagnostic assessments of thermal concepts. For example, Gurcay and Gulbas (2015) developed a three-tier diagnostic test to distinguish correct conceptions from misconceptions about heat, temperature, and internal energy. The present essay-based instrument addresses a complementary question: not only whether a student holds an appropriate concept, but at which stage the concept is successfully or unsuccessfully applied in solving a contextual problem. This distinction is relevant because thermal-physics difficulties can persist even after formal instruction (Leinonen et al., 2013).

The practical contribution is therefore stage-specific feedback. A student who selects the correct heat equation but makes a substitution error requires a different instructional response from a student who cannot identify the relevant physical principle. Likewise, a student who completes the calculation but does not evaluate the plausibility of the result may need explicit practice in checking units, limits, and physical meaning. Structured problem-solving templates have been used for both scaffolding and assessment for this reason (Burkholder et al., 2020).

Several limitations constrain the interpretation of the present statistics. The empirical trial involved only 30 students from one school, so the item parameters and internal-consistency estimate may be sample-specific. The study also used a single rater, which means that inter-rater consistency of the analytic rubric has not yet been established. In addition, corrected item-total correlations and Cronbach's alpha cannot determine whether the proposed five stages are empirically distinct dimensions.

Further validation should therefore use a larger and more diverse sample, multiple trained raters, and analyses that directly examine the structure of the score. Inter-rater reliability should be estimated before the rubric is used for high-stakes decisions. With adequate sample size, factor-analytic or item-response approaches could then test whether the five-stage structure is supported empirically and whether fewer problems can retain sufficient diagnostic information while reducing testing time.

CONCLUSION

This study provides preliminary evidence of validity and reliability for a five-stage, process-oriented assessment of students' problem solving in heat learning. Expert review produced S-CVI/Ave = 0.960 and S-CVI/UA = 0.800, and all 20 sub-items exceeded the study's prespecified corrected item-total correlation criterion. Cronbach's alpha was 0.975, difficulty indices ranged from 0.583 to 0.750, and discrimination indices ranged from 0.406 to 0.750. These results support the continued use of the instrument for exploratory, stage-specific diagnosis, particularly for distinguishing difficulties in problem representation, physical description, planning, calculation, and evaluation. However, the evidence remains preliminary because the trial involved a small sample from a single school, a single rater, and no analysis of the instrument's dimensional structure. Future studies should prioritize multi-site administration, inter-rater reliability, and larger-sample structural validation before the instrument is used for broader or high-stakes assessment.

ACKNOWLEDGEMENTS

The authors express sincere gratitude to the five validators who provided expert judgments on the content validity of the assessment instrument, including their constructive comments and recommendations for improvement. The authors also thank the Grade 12 students who participated in the empirical trial, as well as the school administration and physics teachers who facilitated the data collection process. Appreciation is further extended to the Department of Physics Education, Universitas Pendidikan Indonesia, for supporting the development of this research. The authors declare no conflict of interest regarding the publication of this article.

REFERENCES

Burkholder, E. W., Miles, J. K., Layden, T. J., Wang, K. D., Fritz, A. V., & Wieman, C. E. (2020). Template for teaching and assessment of problem solving in introductory physics. *Physical Review Physics Education Research, 16*(1), 010123. https://doi.org/10.1103/PhysRevPhysEducRes.16.010123

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart, and Winston.

Docktor, J. L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K. A., Mason, A., Ryan, Q. X., & Yang, J. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical Review Physics Education Research, 12*(1), 010130. https://doi.org/10.1103/PhysRevPhysEducRes.12.010130

Gurcay, D., & Gulbas, E. (2015). Development of three-tier heat, temperature, and internal energy diagnostic test. *Research in Science & Technological Education, 33*(2), 197–217. https://doi.org/10.1080/02635143.2015.1018154

Heller, J. I., & Reif, F. (1984). Prescribing effective human problem-solving processes: *Problem description in physics. Cognition and Instruction, 1*(2), 177–216. https://doi.org/10.1207/s1532690xci0102_2

Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics, 60*(7), 627–636. https://doi.org/10.1119/1.17117

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17–24. https://doi.org/10.1037/h0057123

Leinonen, R., Asikainen, M. A., & Hirvonen, P. E. (2013). Overcoming students' misconceptions concerning thermal physics with the aid of hints and peer interaction during a lecture course. *Physical Review Special Topics - Physics Education Research, 9*(2), 020112. https://doi.org/10.1103/PhysRevSTPER.9.020112

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467. https://doi.org/10.1002/nur.20199

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd