GeoScienceEd *7(3) (2026)*

Jurnal Pendidikan, Sains, Geologi, dan Geofisika

http://jpfis.unram.ac.id/index.php/GeoScienceEdu/index

AKM-Based Formative Assessment Instrument for Evaluating Critical Thinking Skills in Thermochemisty Learning

Aryka Khulayda Mardhatilla[1*], Rudiana Agustini[1]

[1] Study Program of Chemistry Edycation, Faculty of Mathematics and Science, Universitas Negeri Surabaya, Indonesia.

DOI: https://doi.org/10.29303/goescienceed.v7i3.2363

Article Info:
Received : May, 28 2026
Revised : June, 16 2026
Accepted : June, 27 2026
Published : July, 27 2026

Correspondence:

Aryka Khulayda Mardhatilla

Phone: +62 857-0787-7080

Abstract: To support the Independent Curriculum and meet 21st century learning expectations, it is critical to design an assessment system that can measure high-level thinking skills. One of the main challenges is the limited availability of formative assessments based on the Minimum Competency Assessment (AKM) that specifically evaluate critical thinking skills in thermochemistry. The goal of this project is to create a formative assessment tool based on the AKM framework that has strong validity and reliability for assessing the thermochemistry-related critical thinking abilities of eleventh-grade high school students. The ADDIE model served as the guiding framework for the development process. Three professional validators assessed the instrument's theoretical validity, while a sample of 34 eleventh-grade students from SMA Hang Tuah 1 Surabaya participated in empirical testing. The findings reveal that the instrument attained an exceptionally high level of theoretical validity, evidenced by content validity of 93.3%, construct validity of 100%, and language validity of 100%. A Cronbach's Alpha coefficient of 0.950 demonstrated the exceptional reliability and empirical validity of all 20 items. The majority of the items on the test were classified as moderately difficult and had a high capacity for discrimination, demonstrating the instrument's good qualitative qualities.

Keywords: Minimum Competency Assessment; Formative Assessment; Critical Thinking Skills; Instrument Development; Thermochemistry.

Citation: Mardhatilla, A. K., & Agustini, R. (2026). AKM-Based Formative Assessment Instrument for Evaluating Critical Thinking Skills in Thermochemisty Learning. *Jurnal Pendidikan, Sains, Geologi, Dan Geofisika (GeoScienceEd Journal)*, *7*(3), 4448–4457. https://doi.org/10.29303/goescienceed.v7i3.2363

Introduction

Evaluation is an essential element in the process of learning, because it contributes to the provision of information about the level of achievement of students' competencies and is used as a basis for pedagogical decision-making (Kusumaningtyas et al., 2020). Regarding the application of the Merdeka Curriculum, measuring learning outcomes is not the only goal of evaluation, it also acts as an instrument to monitor the development of students' competencies in a structured manner in an ongoing manner through a formative assessment approach (Yuni et al., 2025). In this regard, an evaluation instrument is needed that is able to measure competence comprehensively, especially high-level thinking skills as a must in 21st century learning (Kharisma Romadhon & Irfan, 2025).

One of the important aspects of high-level thinking skills that need to be honed intensively is the ability to think critically. This ability allows students to analyze various information, assess the validity of arguments, draw logical conclusions, and solve problems based on evidence that can be accounted for (Ennis, 1996; Facione, 2015). However, a number of actual investigations demonstrate that Indonesian students' critical thinking abilities are still lacking, especially in solving problems in real contexts that require a complex level of reasoning (Muftia et al., 2022). This condition shows the need for an assessment

Email: aryka.22059@mhs.unesa.ac.id

instrument that is capable of measuring and encourage the growth of students' critical thinking abilities.

One of the evaluation models that is in accordance with this imperative is the Minimum Competency Assessment (AKM). AKM is intended to measure reasoning skills through problems in real contexts that require the capacity to methodically assess, analyze, and resolve issues (Sumarni et al., 2023) The characteristics of AKM are in harmony with the indicators of critical thinking abilities, so that they can be used as a basis for developing more meaningful and authentic evaluation development instruments (Damastuti et al., 2022). However, the existence of AKM-based formative assessment instruments specifically designed to measure critical thinking skills in chemistry learning is still very limited (Fahmi & Hernani, 2023). To date, no prior study has developed a formative assessment instrument that specifically combines the AKM framework with critical thinking indicators on thermochemistry content, which marks the gap and the novelty addressed by the present study.

In the context of chemistry learning, the ability to think critically is crucial, considering that students are required to master abstract concepts and the skills to understand phenomena contained in daily life (Suleman et al., 2023). One of the materials that requires this ability is thermochemistry. Thermochemical materials include the concepts of energy change, heat, enthalpy, Hess's law, and bonding energy that demand conceptual analysis and problem-solving skills (Istiqomah & Maemonah, 2021). Various studies show that students still have difficulty understanding these concepts and often experience misconceptions in solving thermochemistry problems (Artini & Wijaya, 2020; ISLAMIYAH, 2025).

Results from in-depth interviews with chemistry teachers and field observations at SMA Hang Tuah 1 Surabaya indicate that current assessment practices remain predominantly summative and rely heavily on question items targeting Low Order Thinking Skills (LOTS). In addition, no AKM based formative assessment instruments are currently available to assess students' critical thinking skills in thermochemistry. This situation underscores the urgent need to develop assessment instruments with high validity and reliability to assess pupils' capacity for critical thought in a more comprehensive and holistic manner.

According to the description given above, the purpose of this study is to develop a formative assessment instrument grounded in the Minimum Competency Assessment (AKM) framework, designed to assess pupils' critical thinking abilities in the eleventh grade on thermochemistry topics. The quality of the instrument is evaluated through theoretical validity, which includes material, construct, and linguistic aspects, as well as empirical validity, encompassing item validity, reliability level, difficulty index, and discriminating power. The resulting instrument is expected to serve as a highly valid and reliable measurement tool that effectively supports the optimal implementation of formative assessment in chemistry learning.

Method

In order to create and offer assessment tools that are especially made to gauge students' critical thinking abilities, this study uses a research and development methodology (Dick, 2001). The ADDIE approach, which consists of five consecutive stages analysis, design, development, implementation, and evaluation, was used in this study (Branch, 2009). The sample consisted of 34 eleventh-grade students selected through purposive sampling, while the instrument used was a formative assessment test complemented by expert validation sheets. Data were analyzed through theoretical validation by three validators and, subsequently, through statistical analysis of item validity, reliability, difficulty index, and discriminating power using SPSS 25.

Development Procedure

The five ADDIE stages are carried out sequentially, with the output of each stage serving as the input for the next: the curriculum and material analysis in the Analysis stage informs the blueprint prepared in the Design stage, which is then realized into the instrument produced in the Development stage, tried out in the Implementation stage, and continuously refined through the Evaluation stage at every step, as illustrated in Figure 1.

In the Analysis stage, the researcher reviewed the applicable curriculum, analyzed thermochemical materials into a concept map, and identified the characteristics of eleventh-grade students, which together served as the basis for determining the assessment needs and critical thinking indicators to be measured. Then, in the Design phase, the researcher formulated the purpose of the assessment, determined the critical thinking skill indicators, prepared a test blueprint containing question grids, decided the form of question items (multiple choice, complex multiple choice, matching, and descriptive), and compiled the scoring guidelines and answer key. Furthermore, at the Development stage, the initial instrument draft was written, reviewed by the supervisor, and validated by three expert validators covering content, construct, and language aspects, after which revisions were made based on the validators' input to produce the final draft.

The Implementation Stage was carried out by administering the validated instrument as a trial to a

sample of 34 students to obtain empirical data on item validity, reliability, difficulty index, and discriminating power. Meanwhile, the Evaluation stage was carried out iteratively at each of the preceding stages, in the form of consultation with the supervisor and validators, to provide feedback and refinement of the processes and products produced, so that the developed instruments have optimal quality and are in accordance with the research objectives.

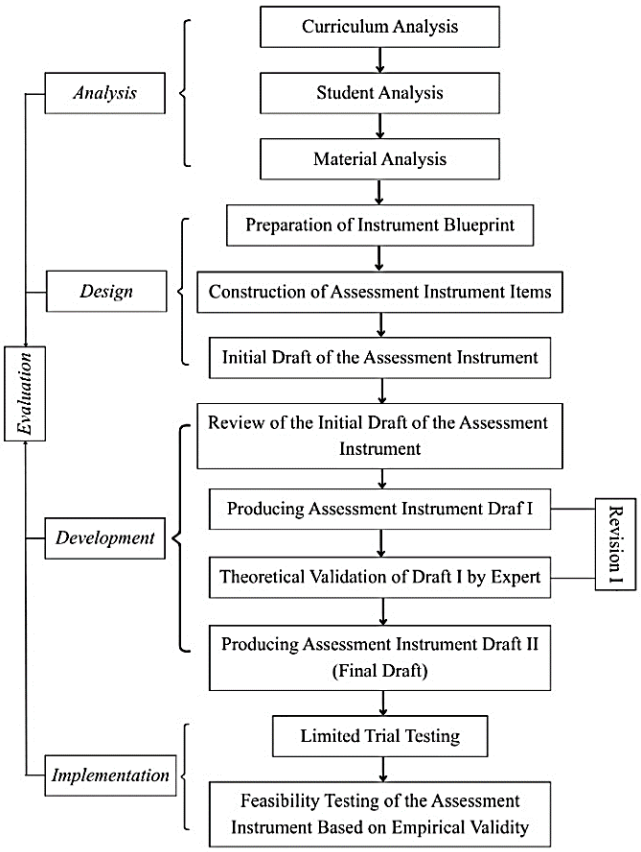

Figure 1. ADDIE development procedure scheme (adapted from Branch, 2009)

Research Place and Subject

The instrument was developed and validated as part of the State University of Surabaya's (Unesa) Undergraduate Chemistry Education Study Program at the Faculty of Mathematics and Natural Sciences. Subsequently, SMA Hang Tuah 1 Surabaya conducted a short study. Purposive sampling was used to pick 34 students from class F.XI 3, Phase F, in the odd semester of the 2025–2026 academic year to participate in the trial.

Data Collection Techniques

Data acqusition was conducted in two phases, namely the instrument validation phase and the instrument trial phase. The research instrument was administered as a paper-based written test with an allotted completion time of 90 minutes.

Data Analysis Techniques

Theoretical Validity

Validity test is the process of evaluating the feasibility of an instrument to be used in research (Sugiyono, 2022). Validity indicates the degree to which the evaluation instrument is capable to measuring the desired level of achievement in accordance with the previously set targets (Zahro, 2020). The validity of the instrument consists of theoretical validity and empirical validity. Theoretical validity is an assessment of instruments based on logical and rational considerations (Arikunto, 2018). Theoretical validity includes aspects of content (material), constructs, and language (Rahayu, 2019). Theoretically valid instruments are a prerequisite for empirical testing (Agustiani et al., 2022). The validity of the theoretical data was obtained through validation sheets given to three validators. The input provided by the validators is utilized as a foundation for improving the tool prior to its implementation of the trial. The assessment uses the Guttman scale with a firm answer choice of "Yes" and "No", as presented in Table 1.

Table 1. Scale Guttman Criteria

Category	Score
Yes	1
No	0

The scores obtained from the three validators are then calculated as a percentage of validity using the formula:

$$\textbf{Percentage}\ (\%) = \frac{\Sigma\ \boldsymbol{Scores\ marked\ "yes"}}{\Sigma\ \boldsymbol{Makximum\ score}}\ \boldsymbol{x}\ \mathbf{100}\%$$

As shown in Table 2, the percentage findings are translated into the validity requirements.

Table 2. Validity Criteria

Validity (%)	Criteria
81.50 ≤ N ≤ 100.0	Very valid
62.75 ≤ N ≤ 81.49	Valid
44.00 ≤ N ≤ 62.75	Fairy valid
25.00 ≤ N ≤ 43.99	Less valid
00.00 ≤ N ≤ 24.99	Invalid

(Riduwan, 2018)

Empirical Validity

The empirical validity as in this study uses statistical analysis to ascertain the degree of suitability of the instrument based on the test outcomes or measurements that have been obtained (Ismail, 2020). The objective of empirical validation is to ascertain the relationship between test results and relevant criteria that are outside the instrument itself, with the requirement that the criteria must be in accordance with the aspect being measured. The implementation of

empirical testing of the validity and reliability of instruments is a really crucial step to guarantee that each question item truly represents the theoretical construct being studied, while supporting the accuracy of the measurement results obtained (Ramadhan, 2024).

Validity of Question Items

The empirical validity of the query item is the value obtained from the calculation of the correlation coefficient between the score of each question item and the overall score. If the score of a question item makes a significant contribution to the total score, then the question item has been declared to have high validity (Ketaren et al., 2024). The more accurate the inferences made according to the test findings, the higher the validity of the tool (Wakhidah & Indana, 2020). The empirical validity analysis of the question items in this study was completed using the Pearson Product Moment Correlation method through SPSS 25 software.

Reliability

Reliability of a test refers to the measuring result's consistency, even when the measurement is carried out under different conditions (Arikunto, 2018). Reliability indicates the degree of reliability of the measurement results and remain accurate across various situations. It is important to understand that although an instrument may be divided into several parts, this does not guarantee that the instrument can be used universally without considering the context and characteristics of the subjects involved (Yusup, 2018). An assessment instrument that is valid will automatically be trustworthy, but a reliable instrument does not always meet the criteria of validity (Ono, 2020).

In this study, reliability was analyzed using an internal consistency approach, namely by testing the instrument once. The intrinsic coherence approach determines the degree each item in the instrument provides consistent or correlated results with each other (Ramadhan, 2024). Reliability tests were performed using Cronbach's Alpha technique through SPSS 25 software. The calculation results are interpreted based on the reliability category in Table 3.

Table 3. Reliability Question Category

Interval	Criteria
0.80 < a ≤ 1.00	Highly reliable
0.60 < a ≤ 0.80	Reliabel
0.40 < a ≤ 0.60	Quite reliable
0.20 < a ≤ 0.40	Less reliable
0.00 < a ≤ 0.20	Unreliable

(Sugiyono, 2022)

Difficulty level

A query is regarded as good when it has an appropriate degree of difficulty, meaning it is neither either simple nor very complex. Students are not challenged by items that are overly simple, while overly difficult items may reduce students' motivation to attempt them (Fatimah & Alfath, 2019). The assessment of item difficulty is based on students' actual performance, not solely on the teacher's assumptions (Magdalena et al., 2021). In arranging the proportion of item difficulty, a balanced ratio of 3:4:3, consisting of 30% easy items, 40% medium items, and 30% difficult items (Imania & Bariah, 2019). The difficulty index is calculated using the formula:

$$\mathbf{P} = \frac{\boldsymbol{Number\ of\ students\ who\ answered\ correct}}{\boldsymbol{Total\ number\ of\ testtakers}}$$

Table 4. shows how the results of difficulty index calculation can be grouped according to the degree of difficulty.

Table 4. Category Difficuly Level Question

Difficulty Index	Category
P < 0.30	Difficult questions
0.30 ≤ P ≤ 0.70	Medium questions
P > 0.70	Easy questions

(Sugiyono, 2022)

Differentiating Power

The distinguishing power of a question item shows how well the item can distinguish between students who have understood the subject and those who have not (Fatimah & Alfath, 2019). This ability is measured using a discriminating power index, which reflects how effectively an item identifies differences in achievement between the upper and lower groups. The validity of a question can also be evaluated from its discriminating power. The item's ability to differentiate between participant's ability levels increases with its discriminating power (Imania & Bariah, 2019). The discriminating power coefficient ranges from -1 to 1. Items with negative coefficient values indicate misleading questions, as the high-ability group answers them incorrectly more often than the low-ability group. Such items should be removed because they do not function properly as measurement tools (Nurhalimah et al., 2022). The differentiating power index is calculated using the formula:

$$\boldsymbol{DB = \frac{BA}{JA} - \frac{BB}{JB} = PA - PB}$$

Description:

D = Discriminating power index

BA= Total number of pupils who provided

accurate answers
BB = Total number of pupils in lower classes who said "yes"
JA = Total top group
JB = Total bottom group
PA = Percentage of the upper group that provided an accurate response
PB = Percentage of the lower group that said "yes"

The results of the scores obtained can be interpreted based on the criteria of differentiating power of the questions in Table 5.

Table 5. Criteria for Differentiating Power of Question

Differentiating Power Index	Category
0,71–1,00	Excellent
0,41–0,70	Good
0,21–0,40	Enough
0,00–0,20	Less

(Arikunto, 2018)

Critical Thinking Skills

This study employs the critical thinking indicators proposed by Facione as the basis for assessing participants' critical thinking skills. The selection of Facione's indicators is grounded in their coherence with the conceptual definition of critical thinking, which highlights an individual's capacity for problem solving. The six indications of critical thinking are interpretation, analysis, evaluation inference, explanation, and self-regulation (Facione, 2015). But, this study utilizes only three of these indicators is analysis, evaluation, and self-regulation on the grounds that they are the most relevant to the education objectives and the attributes of the thermochemistry material. The clasification of pupil's critical thinking abilities according to the assessment findings obtained through the instrument is presented in Table 6.

Table 6. Critical Thinking Skills Category

Value Interval	Category
90–100	Very high
75–89	High
55–74	Moderate
40–54	Low
<40	Very low

(Fitriani & Fauzi, 2024)

Result and Discussion

Analysis

In the analysis stage, a study of the curriculum, student characteristics, and subject matter is carried out as a basis for development instruments. Curriculum Analysis Refers to Phase F Chemistry Learning Outcomes which emphasizes mastery of concepts, science process skills, and critical thinking skills. The analysis of students was carried out on the characteristics of grade XI students of SMA Hang Tuah 1 Surabaya who are in the formal operation stage, where they already have the ability to think logically and systematically. The material analysis is focused on thermochemical materials through the preparation of concept maps to identify problem indicators to be developed. In addition, the evaluation is carried out by verifying the alignment between learning outcomes, student characteristics, materials, indicators, and learning objectives through consultation with supervisors and chemistry educators, so that the instruments developed are in accordance with learning needs and are able to accurately measure students' critical thinking skills.

Design

The Design Stage involves developing an AKM-based formative assessment instrument intended to measure students' critical thinking skills in thermochemistry. The primary activities include constructing a test blueprint aligned with critical thinking indicators at the cognitive levels of analysis, evaluation, and creation. The instrument structure consists of working instructions, student identity information, and 20 items presented in four formats: multiple choice, complex multiple choice, matching, and descriptive questions. The preparation of test items is based on authentic scientific reference sources. Additionally, an answer key, scoring rubric, and time allocation are formulated to produce the initial draft of the instrument.

Subsequently, an evaluation is carried out through discussions with the supervisor to verify the alignment of the blueprint, indicators, item types, and scoring guidelines with AKM characteristics and the objectives of measuring critical thinking skills. The evaluation outcomes serve as the basis for refining indicators, revising item wording, improving scoring rubrics, and enhancing contextual stimuli, thereby increasing the instrument's validity and preparing it for further development in the next stage.

Development

At the Development stage, the initial draft of the instrument was reviewed by the supervisor and subsequently revised into Draft 1. Draft 1 was then subjected to theoretical validation by two Chemistry Education lecturers and one high school chemistry teacher, who assessed its validity in terms of content, construct, and language. The validation results indicated that the instrument met the established validity criteria

and included several recommendations for improvement, which were incorporated into revisions that produced Draft 2 as the final version of the instrument. The results of the theoretical validation conducted by the three validators are presented in table below.

Table 7. Material validity results

Validator	Score of Material Aspect (%)
Validator 1	87,5
Validator 2	100
Validator 3	92,5
Average	93,3
Interpretation	Very valid

Table 8. Construct validity results

Validator	Score of Construct Aspect (%)
Validator 1	100
Validator 2	100
Validator 3	100
Average	100
Interpretation	Very valid

Tabel 9. Language validity results

Validator	Score of Language Aspect (%)
Validator 1	100
Validator 2	100
Validator 3	100
Average	100
Interpretation	Very valid

Evaluation at this stage is carried out based on the results of validation and validator input. The results of the evaluation were used to improve the instruments so that they obtained instruments that were theoretically valid and ready to be tested on 34 students in grade XI of SMA Hang Tuah 1 Surabaya.

Implementation

At the Implementation stage, the validated assessment instrument was administered on a limited scale to 34 eleventh-grade students at SMA Hang Tuah 1 Surabaya. The purpose of this trial was to obtain empirical data to evaluate the quality of the instrument, including item validity, reliability level, difficulty index, and discriminating power. The instrument used in the trial was the final draft produced after revisions based on expert validation. The results of the trial and the subsequent analysis of the instrument's quality are presented in the following discussion section.

Based on Table 10, all 20 items were declared to be emirically valid because the calculated r-values ranged from 0.443 to 0.920, exceeding the r-table value of 0.339. These results show that each item has a significant correlation with the total score and is able to measure the construct in question, i.e. students' critical thinking skills in thermochemistry. A high item-total correlation value indicates that the developed items effectively distinguish students with different levels of critical thinking ability. Some items showed very strong correlation, such as item 19 (r = 0.920) and item 20 (r = 0.866), which showed their strong contribution to the overall measurement. Although item 10 obtained the lowest correlation coefficient (r = 0.443), the value still exceeded the minimum validity criteria, which means it is still suitable for use.

Tabel 10. Empirical Validity Results of each Question Item

Question Item	Calculated r value	Interpretation
1	0,832	Valid
2	0,622	Valid
3	0,625	Valid
4	0,760	Valid
5	0,829	Valid
6	0,735	Valid
7	0,609	Valid
8	0,758	Valid
9	0,812	Valid
10	0,443	Valid
11	0,657	Valid
12	0,768	Valid
13	0,812	Valid
14	0,536	Valid
15	0,777	Valid
16	0,589	Valid
17	0,691	Valid
18	0,827	Valid
19	0,920	Valid
20	0,866	Valid

Tabel 11. Reliability statistic

Cronbach's Alpha	N of Items
0.950	20

Based on Table 11, the reliability test yielded a Cronbach's Alpha value of 0.950 for a 20-item instrument. This value indicates a very high level of reliability, as it exceeds the generally accepted minimum threshold for educational research. The results showed that the items consistently measured the same construct, i.e. students' critical thinking skills in thermochemistry. Cronbach's high Alpha values indicate strong internal consistency among the items, which indicates that the instrument produces stable and reliable measurement results (Sobiri, 2026). Therefore, the developed instrument can be considered reliable and suitable for assessing students' critical thinking skills in

thermochemistry learning. In addition, these findings support the results of empirical validity, which confirm that the instrument is of good quality and can be applied in educational assessment and research.

Tabel 12. Difficulty Level Results for each Question Item

Question Item	Difficulty Level Value	Interpretation
1	0,60	Medium
2	0,56	Medium
3	0,59	Medium
4	0,77	Easy
5	0,38	Medium
6	0,58	Medium
7	0,26	Difficult
8	0,74	Easy
9	0,58	Medium
10	0,28	Difficult
11	0,44	Medium
12	0,88	Easy
13	0,63	Medium
14	0,43	Medium
15	0,58	Medium
16	0,57	Medium
17	0,32	Medium
18	0,57	Medium
19	0,63	Medium
20	0,59	Medium

Based on Table 12, the difficulty level analysis shows that the instrument consists of 3 easy questions (points 4, 8, and 12), 16 medium questions, and 2 hard question (point 7 and 10). The difficulty index value ranges from 0.26 to 0.88, indicating variations in the item's difficulty. The dominance of intermediate category items indicates that the instrument has a balanced level of difficulty and is suitable for assessing students with different levels of understanding. Items with a moderate level of difficulty are generally considered the most effective for measuring learning outcomes because they can distinguish students who have mastered the material from those who have not. Meanwhile, the existence of easy and difficult items complements the instrument by allowing the assessment of low-performing and high-performing students.

Based on Table 13, the analysis of the discriminating power reveals that 11 items are categorized as good, 8 items as adequate, and 1 item as bad. The value of the discrimination index ranges from 0.14 to 0.70. These findings suggest that most of the items developed have adequate to good ability to distinguish between students with high and low levels of critical thinking skills. Items with good discriminating power, such as items 1, 2, 6, 13, and 15, effectively distinguish students who have mastered the material from those who have not. Meanwhile, items classified as sufficient still meet acceptable standards and can contribute to the overall assessment. In contrast, item 10 showed a low discrimination index (0.14), which suggests that the item was less effective in differentiating students based on their ability level and may require revision.

Tabel 13. The Differentiating Power of Each Question Item

Question Item	Differentation Values	Interpretation
1	0,65	Good
2	0,60	Good
3	0,45	Good
4	0,48	Good
5	0,48	Good
6	0,60	Good
7	0,24	Enough
8	0,43	Good
9	0,52	Good
10	0,14	Less
11	0,28	Enough
12	0,26	Enough
13	0,70	Good
14	0,23	Enough
15	0,60	Good
16	0,38	Enough
17	0,33	Enough
18	0,34	Enough
19	0,34	Enough
20	0,43	Good

Tabel 14. Students' Level of Critical Thinking Ability in Thermochemistry Materials

Name	Value	Category Critical Thinking Learners
AND	43	Low
AKW	40	Low
AAF	35	Very low
AP	47	Low
BAP	41	Low
CPP	52	Low
EAC	74	Moderate
FR	31	Low
FKA	47	Low
GMG	76	High
HAS	51	Low
IDM	46	Low
KAA	48	Low
LMOM	35	Very low
LCP	59	Moderate
MDP	72	Moderate
MAC	56	Moderate
MRA	38	Very low
MFAS	31	Very low

Name	Value	Category Critical Thinking Learners
MGSPS	36	Very low
MLP	66	Moderate
MR	39	Very low
MKK	76	High
MNI	62	Moderate
NPM	85	High
PSA	48	Low
RGPH	60	Moderate
RAP	75	High
RAS	40	Low
RK	71	Moderate
RS	79	High
SAA	53	Low
SAM	78	High
SNPP	85	High

Based on measurement of critical thinking skills on 34 students, the distribution was obtained as many as 7 students (20.59%) were in the very low category, 14 students (41.18%) in the low category, 7 students (20.59%) in the medium category, and 6 students (17.65%) in the high category, while no students reached the very high category. These results show that students' critical thinking skills are still dominated by the low and very low categories with a combined percentage of 61.77%. These findings indicate the need for learning efforts that focus more on developing critical thinking skills so that students are able to analyze, interpret, and solve thermochemical problems more effectively. Overall, the instruments developed have functioned well as a tool to measure students' critical thinking skills on thermochemical materials. This instrument is able to group students into various levels of ability clearly and systematically, so it is feasible to use to identify students' critical thinking skill levels. However, given that most students are still in the low and very low categories, further research is needed to develop and implement more effective learning strategies in improving students' high-level thinking skills.

Conclusion

This study developed a formative assessment instrument based on the Minimum Competency Assessment (AKM) framework to evaluate the critical thinking skills of eleventh-grade high school students in thermochemistry. The instrument demonstrated excellent theoretical validity, achieving a score of 93.3% for content validity, 100% for construct validity, and 100% for language validity, all of which fall into the category of highly valid. Further empirical testing confirmed that all twenty items were valid and showed very high reliability, as indicated by the Alpha Cronbach coefficient of 0.950. In addition, the instrument exhibits satisfactory psychometric properties, with most items classified as having a moderate level of difficulty and good to adequate discriminating power. The results of the assessment revealed that 20.59% of students were categorized as having very low critical thinking skills, 41.18% low, 20.59% moderate, and 17.65% high, while none of the students reached the very high category. Overall, 61.77% of students fell into the low and very low categories, indicating that their critical thinking skills in thermochemistry are still relatively underdeveloped. These findings point to the need for more effective teaching approaches that aim to encourage higher-level thinking skills. Therefore, although the AKM-based formative assessment instruments developed in this study have been proven to be valid and reliable, further research is recommended to design and implement learning strategies that can improve students' critical thinking skills. As a result, the instrument can not only serve as an assessment tool but also as a basis for improving the quality of teaching and learning.

Acknowledgements

Throughout this study process, the writer realizes that the implementation of research will not be successful without the backing of many parties. Consequently, the author expresses his sincere gratitude to both parents who always provide prayers for blessings, help, amenities that are required and assistance that never stops during the course of inestigation. The author also thanked the supervisor for the guidance, direction, motivation, and advice that was very valuable in the implementation of this research. The writer wishes to convey the sincere appreciation to the validators who have donated their time, energy, and expertise to provide valuable assessments, inputs, suggestions, and feedback that are very useful in the process of validating and improving the developed product. The author expresses high appreciation to the principals, teachers, and students who have been willing to participate in the data collection process. In addition, the author also expressed the appreciation to colleagues who have supplied moral support and enthusiasm during the research process. Success in completing this research and development will not be realized without the help, cooperation, and support of all parties.

References

Agustiani, N., Setiani, A., & Lukman, H. S. (2022). Pengembangan Instrumen Tes PLSV Berdasarkan Indikator Berpikir Kritis dan Pemecahan Masalah. Jambura Journal of Mathematics Education, 3(2). https://doi.org/10.34312/jmathedu.v3i2.15837

Arikunto. (2018). Dasar-Dasar Evaluasi Pendidikan (3rd ed.).

Artini, N. P. J., & Wijaya, I. K. W. B. (2020). Strategi pengembangan literasi kimia bagi siswa SMP. Jurnal Ilmiah Pendidikan Citra Bakti, 7(2). https://doi.org/10.38048/jipcb.v7i2.97

Branch, R. M. (2009). Instructional Design: The ADDIE Approach. Springer US. https://doi.org/10.1007/978-0-387-09506-6

Damastuti, C. S., Rinjani, A. M., & Vitaningksih, M. (2022). Strategi Guru SD dalam Menghadapi Assesmen Kompetensi Minimum. Jurnal Pedagogik Pendidikan Dasar, 9(1), 11–16. https://doi.org/10.17509/jppd.v9i1.43168

Dick, W. , & C. L. (2001). The systematic design of instruction (5th ed.). Allyn & Bacon.

Ennis, R. H. (1996). Critical thinking. Prentice Hall.

Facione, P. A. (2015). Critical thinking: What it is and why it counts. Insight Assessment, 5(1), 1–30.

Fahmi, an, & Hernani, dan. (2023). Pengembangan Asesmen Formatif Berbasis Problem Based Learning untuk Meningkatkan Keterampilan Berpikir Kritis: Need Assessment Study. Orbital: Jurnal Pendidikan Kimia, 7(2).

Fatimah, L. U., & Alfath, K. (2019). Analisis Kesukaran Soal, Daya Pembeda Dan Fungsi Distraktor. Al-Manar, 8(2). https://doi.org/10.36668/jal.v8i2.115

Fitriani, E., & Fauzi, A. (2024). Meningkatkan Keterampilan Berpikir Kritis Dengan Model PBL-RME Pada Pembelajaran Matematika Sekolah Dasar. Pendas : Jurnal Ilmiah Pendidikan Dasar, 9(04), 210–221. https://doi.org/10.23969/jp.v9i04.20109

Imania, K. A., & Bariah, S. K. (2019). Rancangan Pengembangan Instrumen Penilaian Pembelajaran Berbasis Daring. Jurnal Petik, 5(1). https://doi.org/10.31980/jpetik.v5i1.445

Islamiyah, A. (2025). Analisis Miskonsepsi Siswa Menggunakan Instrumen Tes Diagnostik Five-Tier Padamateri Termokimia . Universitas Islam Negeri Sultan Syarif Kasim Riau.

Ismail, M. I. (2020). Evaluasi Pembelajaran: Konsep Dasar, Prinsip, Teknik dan Prosedur. Rajawali Pers, 1.

Istiqomah, N., & Maemonah, M. (2021). Konsep Dasar Teori Perkembangan Kognitif Pada Anak Usia Dini Menurut Jean Piaget. Khazanah Pendidikan, 15(2). https://doi.org/10.30595/jkp.v15i2.10974

Ketaren, M. A., Girsang, K., Manurung, M., & Ginting, E. R. B. (2024). Uji Validitas Dan Uji Daya Beda Soal Buatan Pilihan Ganda Dengan Tes Sumatif Siswa Kelas IV UPT SD Negeri 065013 Medan. Jurnal Pengabdian Masyarakat Bangsa, 1(11). https://doi.org/10.59837/jpmba.v1i11.706

Kharisma Romadhon, & Irfan, I. (2025). Analisis Kompetensi Guru terhadap Penyusunan Perangkat Pembelajaran pada Kurikulum Merdeka di Sekolah Dasar. Scholaria: Jurnal Pendidikan Dan Kebudayaan, 15(2). https://doi.org/10.24246/j.js.2025.v15.i2.p111-123

Kusumaningtyas, P., Oktafiani, R., Nurhadi, M., Sulistyaningwarni, S., Studi Sarjana Pendidikan Kimia Universitas Mulawarman, P., & Menengah Atas Negeri, S. (2020). Pengaruh Isu Sosiosaintifik Dalam Model Discovery Learning Terhadap Keterampilan Berpikir Kritis Siswa Pada Materi Asam Basa. Orbital: Jurnal Pendidikan Kimia, 4(1).

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas Iii Sdn Karet 1 Sepatan. Bintang : Jurnal Pendidikan dan Sains (Vol. 3, Number 2). https://ejournal.stitpn.ac.id/index.php/bintang

Muderawan. (2019). Analisis Faktor-faktor Penyebab Kesulitan Belajar Siswa pada Materi Kelarutan dan Hasil Kali Kelarutan.

Muftia, N., Aziz, M. A., & Indriani, E. (2022). Meta Analisis Pengaruh Model Pembelajaran Discovery Learning Terhadap Kemampuan Berfikir Kritis Siswa Sekolah Dasar. Jurnal Pendidikan Dan Konseling, 4(20).

Nurhalimah, S., Hidayati, Y., Rosidi, I., Wiwin, D., & Hadi, P. (2022). HUBUNGAN ANTARA VALIDITAS ITEM DENGAN DAYA PEMBEDA DAN TINGKAT KESUKARAN SOAL PILIHAN GANDA PAS. In Jurnal Natural Science Educational Research (Vol. 4, Number 3).

Ono, S. (2020). Uji Validitas dan Reliabilitas Alat Ukur SG Posture Evaluation. Jurnal Keterapian Fisik, 5(1). https://doi.org/10.37341/jkf.v5i1.167

Rahayu, C. (2019). Validitas Perangkat Pembelajaran Fisika Sma Berbasis Model Pembelajaran Generatif Dengan Pendekatan Open-Ended Problem Untuk Menstimulus Keterampilan Berpikir Kritis Peserta Didik. Jurnal Pendidikan Fisika, 7(1). http://journal.uin-alauddin.ac.id/indeks.php/PendidikanFisika

Ramadhan, M. F. , S. R. A. , & A. M. W. (2024). Validitas and reliabilitas. Journal on Education, 6(2), 10967–10975.

Riduwan. (2018). Dasar-Dasar Statistika. Alfabeta.

Sugiyono. (2022). Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif, dan R&D). CV Alfabeta.

Suleman, N., Lukum, A., Rauf, N., Paputungan, M., Alio, L., & Sukamto, K. (2023). Identifikasi Pemahaman Konsep Siswa pada Materi Termokimia Menggunakan Tes Diagnostik Three-Tier Multiple Choice. Jambura Journal of Educational Chemistry, 5(2). https://doi.org/10.34312/jjec.v5i2.13255

Sumarni, S., Prayitno, A. T., Syafari, R., Basir, M. A., Febrianti, D., & Putri, A. (2023). Instrumen Tes Kemampuan Literasi Numerasi Berbasis Budaya Lokal Kabupaten Kuningan Jawa Barat. Aksioma: Jurnal Program Studi Pendidikan Matematika, 12(1), 789. https://doi.org/10.24127/ajpm.v12i1.6439

Wakhidah, T. N., & Indana, S. (2020). Validitas dan Reliabilitas Tes Elektronik (E-Test) Materi Ekosistem untuk Mengukur Kemampuan Berpikir Kritis Siswa Kelas X SMA. Berkala Ilmiah Pendidikan Biologi (BioEdu), 10(1). https://doi.org/10.26740/bioedu.v10n1.p171-176

Yuni, R., Rahman, R., Juita, R., Dasar, P., Adzkia, U., & Profesi Guru, P. (2025). Penerapan Asesmen Formatif Dan Sumatif Dalam Kurikulum Merdeka.

Yusup, F. (2018). Uji Validitas dan Reliabilitas Instrumen Penelitian Kuantitatif. Jurnal Tarbiyah : Jurnal Ilmiah Kependidikan, 7(1). https://doi.org/10.18592/tarbiyah.v7i1.2100

Zahro, U. R. (2020). Pengembangan Instrumen Tes Untuk Mengukur Kemampuan Literasi Sains Siswa SMP Pada Tema Pemanasan Global. Lib.Unnes.Ac.Id, 04(02).