Physics Communication
Volume 9 Issue 2 Year 2025 Pages 100-113

P-ISSN 2528-5971 | E-ISSN 2528-598X
https://journal.unnes.ac.id/journals/pc
https://doi.org/10.15294/pc.v9i2.30530

Development of a Critical Thinking Skills Assessment Instrument on Dynamic Electricity for High School Students

Suratni Agustinih*, Wasino, Agus Yuwono

Educational Research and Evaluation, Graduate School, Universitas Negeri Semarang, Indonesia

*Correspondence to: *agustinihsuratni@gmail.com*

Abstract: The development of critical thinking skills is a curriculum mandate in Indonesia to prepare students for global challenges, yet existing assessments often fail to effectively measure these higher-order competencies. This study aims to analyze the developmental needs, characteristics, feasibility of an assessment instrument, and the effectiveness of its use in evaluating the critical thinking skills (CTS) of high school students on the topic of dynamic electricity. This is a research and development (R&D) study that employs the Four-D Model (define, design, develop, disseminate), as developed by Thiagarajan. The sample consisted of 30 students for a small-scale trial and 150 students for a large-scale trial. Data were collected through interviews, tests, questionnaires, and documentation. The data obtained from these methods were then analyzed using the Rasch Model to evaluate the psychometric properties of the instrument. The instrument was characterized by high-cognitive-level multiple-choice questions administered via the Google Forms platform integrated with AutoProctor. The research findings revealed the need to develop a critical thinking skills assessment instrument that meets the requirements of teachers, students, and conceptual frameworks. The instrument's validity was established through two methods. First, an expert validation process yielded an average Aiken's V coefficient of 0.93. Second, data from empirical trials analyzed with the Rasch Model showed that all items met the ZSTD fit criteria for validity (-2.0 < ZSTD < +2.0). Item reliability was 0.74 and 0.92, with Cronbach's Alpha values of 0.84 and 0.87 (reliable). The item difficulty level was evenly distributed, the item discrimination index was 3.44 (good), and the instrument was free from bias. The average score across the 6 CTS indicators was 79.82, indicating that the assessment instrument for critical thinking skills on dynamic electricity is both feasible and effective for use.

Keywords: 4D model, Critical Thinking, Psychometric Properties, Instrument Development, Rasch Model.

Submitted: 2025-07-18. **Revised:** 2025-08-11. **Accepted:** 2025-08-21.

Introduction

The Indonesian curriculum, through regulations like the Minister of Education and Culture Regulation No. 20 of 2016 and frameworks such as the Merdeka Curriculum, mandates the development of critical thinking skills for all students (Kurikulum Pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, Dan Jenjang Pendidikan Menengah, 2024; Standar Kompetensi Lulusan, 2016). These skills are essential for producing a generation capable of competing globally and adapting to scientific advancements. Assessment is a vital component in this process, serving as the primary tool for teachers to measure the effectiveness of learning and ensure students achieve these higher-order cognitive competencies (Arifin, 2019).

However, a significant gap exists between curricular goals and classroom practice. The current status of test instruments in many schools is inadequate for measuring critical thinking, as teachers rarely implement dedicated critical thinking assessments due to various constraints, such as time and a lack of ready-to-use, validated tools (Miftahussa'adiah et al., 2020). Most assessments are still based on Lower-Order Thinking Skills (LOTS) items, which primarily test recall and basic understanding rather than analysis, evaluation, and creation. This shortcoming leads to a critical research gap: the absence of a feasible and reliable instrument designed specifically to measure high school students' critical thinking skills, which contributes to the persistently low levels of these skills among students (Asriadi & Hadi, 2021; Desilva et al., 2020; Hasan et al., 2020). Furthermore, teachers rarely implement dedicated critical thinking assessments due to various constraints, such as time and a lack of ready-to-use, validated tools (Miftahussa'adiah et al.,

2020). This shortcoming leads to a critical research gap: the absence of a feasible and reliable instrument designed specifically to measure high school students' critical thinking skills in subjects like physics, which contributes to the persistently low levels of these skills among students (Asriadi & Hadi, 2021; Desilva et al., 2020; Hasan et al., 2020).

To address this gap, this study focuses on the development of a specific critical thinking test instrument. The development process follows the rigorous Four-D (4D) model, a systematic framework that includes the stages of Define, Design, Develop, and Disseminate to ensure the final product is both valid and effective (Sihombing et al., 2024). The instrument developed is a set of high-cognitive-level multiple-choice questions. This format was chosen because it allows for objective scoring and efficient administration to large groups of students, while still being designed to trigger higher-order thinking processes through complex, contextual scenarios. To further enhance its utility and address modern challenges like academic integrity, the instrument is administered via the Google Forms platform integrated with AutoProctor. This technological integration makes the assessment more accessible and helps minimize opportunities for cheating during online administration (Setiawan, 2024).

Therefore, this study aims to produce a feasible, valid, reliable, and effective instrument to comprehensively measure the critical thinking skills of high school students on the topic of dynamic electricity. To achieve this goal, a systematic Research and Development (R&D) approach using the Four-D model was employed to ensure a structured process from needs analysis to final product dissemination. The quality of the instrument was rigorously evaluated to confirm its feasibility, with validity established through expert judgment and empirical data analysis using the Rasch Model, and reliability confirmed through Cronbach's Alpha calculations. The resulting product is a digital assessment instrument administered via Google Forms and integrated with AutoProctor, designed specifically to address the lack of tools for measuring higher-order thinking skills in the context of physics education.

Methods

Research Design

This study employed a Research and Development (R&D) design using the Four-D (4D) Model, which consists of four stages: Define, Design, Develop, and Disseminate. The systematic workflow of this entire research process is visually outlined in Figure 1, which illustrates the research flow based on the Four-D Model. The R&D method was chosen because it is conducted systematically to develop and test the effectiveness of a new product or innovation (Sugiyono, 2017; Waruwu, 2024). The 4D model is a framework frequently used in educational research to create and validate educational products (Sihombing et al., 2024).

Research Subjects

The research subjects were selected from several public high schools in Semarang, Indonesia. The study was conducted in two phases:

a. A small-scale trial involving 30 Grade XII (Phase F) students from SMAN 5 Semarang.
b. A large-scale trial involving 150 Phase F students from three different schools: SMAN 5 Semarang, SMAN 2 Semarang, and SMAN 3 Semarang.

Data Collection

Data were collected using a multi-method approach to ensure comprehensiveness:

a. Interviews: Conducted with a physics teacher to gather information for the needs analysis and with students to confirm their thought processes behind test answers.
b. Tests: A multiple-choice test was the primary instrument used to measure students' understanding of dynamic electricity and their critical thinking skills.
c. Questionnaires: Used during the small-scale trial to obtain feedback from students regarding the readability of the developed instrument (Arifin, 2019).
d. Documentation: Utilized to gather administrative data, such as student names and their existing test

scores.

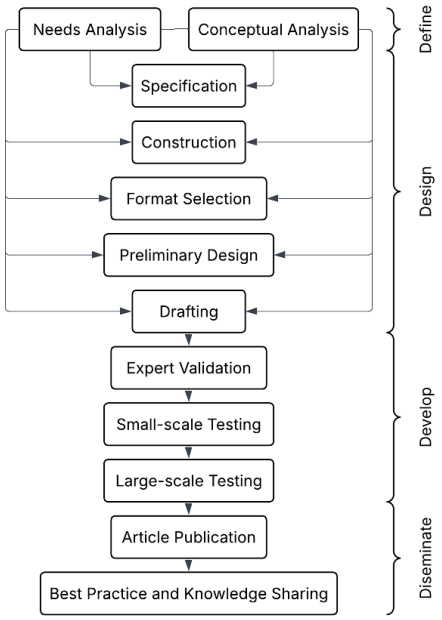

Figure 1. Research flow with 4D model

Data Analysis

The collected data were analyzed using several techniques to determine the instrument's feasibility and effectiveness.

Validity Analysis

Validity was assessed through expert judgment and empirical trials. For expert judgment, Aiken's V formula was used to calculate the content validity coefficient, as shown in Equation (1):

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

Where 's' represents the score given by each expert, which is determined by subtracting 'lo' (the lowest possible rating) from 'r' (the rating given by an expert). The total number of experts providing ratings is represented by 'n', and 'c' represents the number of categories in the rating scale being used.

For empirical data, analysis was conducted using the Rasch Model, with item validity determined by the following criteria (Table 1).

Table 1. Validity Criteria

Criteria	Interval	Conclusion
MNSQ	0.5 < MNSQ < 1.5	Accepted
ZSTD	-2 < ZSTD < 2	Accepted
Pt Measure Corr	0.4 < Pt. Measure Corr < 0.85	Accepted

Reliability, Difficulty, and Discrimination Analysis

The reliability, difficulty level, and discrimination power of the items were also analyzed using the Rasch Model, based on the following criteria (Table 2, Table 3, and Table 4, respectively).

Item Bias and Effectiveness Analysis

An item was identified as biased if it met one of the following criteria: 1) a probability value < 5%; 2) a Mantel-Haenszel value > 1.5; or 3) a DIF contrast value > 0.64. The effectiveness of the instrument was determined by calculating the percentage of student achievement using the formula P = (f/N) x 100%, with

the results categorized based on established criteria.

Table 2. Reliability Criteria

Score Interval	Interpretation
< 0.67	weak
0.67 – 0.80	adequate
0.81 – 0.90	good
> 0.90	excellent

Table 3. Item Difficulty Criteria

Item Category	Logit Value
very easy	≤ -2.0
easy	-2.0 to -0.5
medium	-0.5 to +0.5
hard	+0.5 to +2.0
very hard	≥ +2.0

Table 4. Item Discrimination Criteria

Interpretation	Index Value
Poor	≤ 2
Adequate	2 - 3
Good	3 - 4
Very Good	4 - 5
Excellent	> 5

Results and Discussion

Instrument Development Needs

The research identified several needs for developing the assessment instrument, including student needs, teacher needs, and concept-related needs.

A Student Needs Analysis was conducted using a student needs questionnaire that consisted of seven aspects: subject matter, paper-based assessment, digital assessment, digital devices, question variety, Google Forms assessment, and opening other browsers during assessment. To ground the instrument's development in authentic classroom needs, a questionnaire was administered to 40 students at SMAN 5 Semarang. The data in Table 1 shows a strong student demand for digital assessment tools (95%) and varied questions (90%). However, direct observation confirmed that classroom practice remains conventional, as the assessment process for dynamic electricity material at SMAN 5 Semarang is still conducted using traditional paper-based tests.

Table 5. Student Needs Analysis

Aspect	Percentage (%)
dynamic electricity material	100
use of paper in assessment	100
need for digital assessment tools	95
availability of digital devices	100
variety in question difficulty	90
experience with google forms	100
accessing other browsers during the assessment	85

A majority (95%) of students required engaging and enjoyable digital assessment tools because they could immediately see their scores and the correct answers when using digital-based assessments. The questionnaire indicated that 100% of students had devices to access Google Forms, and all students had already used Google Forms, implying its use in other subjects, as 100% reported using paper for the dynamic electricity assessment. Most students (90%) desired varied questions, which aligns with interview findings that varied questions can identify the diverse ability levels of students. As 85% of students admitted to opening other browsers while using Google Forms, a control mechanism like the AutoProctor add-on is necessary for teachers to monitor students. With AutoProctor, teachers can monitor the sound, browser, and camera used by students. Thus, a digital-based assessment tool can be utilized by teachers to make learning assessments more engaging and enjoyable.

A teacher needs analysis was conducted through interviews and observations. In an interview with a physics teacher at SMAN 5 Semarang. The teacher stated:

> "Jujur, saya jarang sekali membuat soal sendiri. Biasanya saya mengambil dari buku paket atau internet karena lebih praktis. Waktu untuk membuat soal HOTS yang benar-benar mengukur berpikir kritis itu terbatas sekali." (Honestly, I rarely create my own questions. I usually take them from textbooks or the internet because it's more practical. The time to create HOTS questions that truly measure critical thinking is very limited).

This feedback confirms that the existing assessment ecosystem fails to adequately support the measurement of critical thinking, creating a clear gap that this research aimed to fill.

Observations in the classroom at SMA Negeri 5 Semarang revealed that the physics teacher primarily employed the lecture method. This was due to time limitations, the need to cover the curriculum, and inadequate facilities. This teacher-centered approach led to a lack of active student involvement, which hindered the development of critical thinking skills. Most students were unable to connect what they learned with its practical application. Consequently, students tended to use only a fraction of their potential and did not think critically. The teacher also had not implemented a critical thinking skills assessment, which was confirmed by a needs questionnaire distributed to 40 students at SMAN 5 Semarang. Students required a Google Forms-based critical thinking assessment instrument for the topic of dynamic electricity. Other findings showed that some schools had not utilized digital-based assessment, with some teachers confirming they had never used it at all. Therefore, the interview and questionnaire results indicate a clear need for the development of a Google Forms-based assessment instrument to evaluate critical thinking skills in dynamic electricity for high school students.

Concept Analysis involved selecting critical thinking aspects aligned with the curriculum. The concepts in dynamic electricity include Ohm's law, series and parallel circuits, Kirchhoff's laws, and electrical power calculations. These concepts can be applied to everyday problems that require critical thinking, such as analyzing household circuits, interpreting electricity usage data, or evaluating the efficiency of electronic devices. Therefore, the items must be designed to encourage students to interpret data, construct logical arguments, and choose the most appropriate solution based on scientific reasoning (Facione, 2015). In developing multiple-choice questions, stimuli can include voltage-time graphs, circuit diagrams, measurement data tables, or everyday scenarios. According to Zoller (2001), using authentic contexts can increase student engagement and make the thinking process more natural. This is also consistent with the contextual teaching and learning approach widely applied in the Merdeka Curriculum. The critical thinking indicators used align with Facione (2015), namely interpretation, analysis, evaluation, inference, and explanation. The conceptual analysis needs for this instrument include: 1) Alignment of concepts with high cognitive levels; 2) Contextualization and authentic stimuli; 3) Structure of critical and high-quality items; 4) Instrument validation and reliability; 5) An item analysis rubric for critical evaluation; and 6) Integration with the curriculum and the Pancasila Student Profile.

Instrument Characteristics

The critical thinking skills assessment instrument developed in this study has several key

characteristics that define its structure, content, and delivery method. The specifications of the developed product are shown clearly in the following points:

a. Instrument type: An assessment tool designed to measure the critical thinking skills of high school students.
b. Question format: The instrument consists of 20 multiple-choice items. Each item includes a single correct answer and four distractors.
c. Cognitive level: All items are designed to assess Higher-Order Thinking Skills (HOTS) rather than simple recall.
d. Content focus: The subject matter is focused on the topic of Dynamic Electricity for Grade XII high school students.
e. Theoretical framework: The instrument is built upon the six core critical thinking skills as defined by Facione: interpretation, analysis, evaluation, inference, explanation, and self-regulation.
f. Delivery platform: The assessment is administered online using the Google Forms platform, which is integrated with the AutoProctor application to ensure academic integrity.
g. Stimulus variety: The questions utilize diverse stimuli to present contextual problems, including discourse, images, graphs, and tables.
h. Feedback mechanism: Upon completion, the system provides students with immediate feedback, including their final score and a review of the correct answers.

The developed assessment instrument, in the form of multiple-choice questions, measures students' critical thinking skills on dynamic electricity using the Google Forms platform integrated with the AutoProctor automatic proctoring system, representing a strategic innovation in technology-based education. This innovation demonstrates a synergy between critical pedagogy and the use of digital technology to ensure academic integrity and assessment effectiveness. This instrument is not merely an evaluation tool but also a vehicle for developing students' higher-order thinking skills.

The critical thinking skills assessment instrument used in this study was developed based on Facione's (2015) critical thinking skills framework, which comprises six main sub-skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. The instrument consists of multiple-choice questions with stimuli presented as everyday events related to electrical phenomena. Each item is structured with a contextual scenario that challenges students to analyze, evaluate, and make informed decisions based on the data or statements provided. Thus, the instrument assesses high-level cognitive abilities rather than recall.

The integration of the 6 CTS indicators into the dynamic electricity material is broken down as follows:

a. Interpretation
    1. Presenting graphical data results in a conceptual narrative,
    2. Concluding that an object follows Ohm's law,
    3. Giving reasons for voltage and current comparisons.
b. Analysis
    1. Redrawing the circuit correctly,
    2. Explaining the current division,
    3. Providing logical reasons based on Kirchhoff's law or Ohm's law.
c. Evaluation
    1. Stating disagreement,
    2. Explaining the relationship in a voltage equation,
    3. Example calculation or analogy.
d. Inference
    1. Calculating voltage, current, or resistance in a sample circuit diagram,

2. Concluding that an electrical component follows existing electrical laws.

e. Explanation
   1. Explaining the voltage of lamps in series and parallel circuits,
   2. Explaining the relationship between the concepts of voltage, current, and resistance.

f. Self-Regulation
   1. Clearly stating the complex parts
   2. Stating learning strategies or improvements
   3. Formulating critical or meaningful questions (1 point)

The Google Forms platform provides easy access for both teachers and students. This instrument can be accessed anytime and anywhere, as long as an internet connection is available. This facilitates online assessment without logistical hurdles such as printing, physical distribution of questions, or manual collection of answer sheets. The key advantage of this instrument lies in its integration with AutoProctor, an AI-based online exam proctoring application. AutoProctor enables teachers to monitor students during tests via their device's camera and microphone, detecting attempts to open new tabs, use other applications, and even suspicious movements that could indicate cheating (Sivakumar et al., 2024). Thus, students' academic integrity can be optimally maintained. The multiple-choice system, integrated with Google Forms, enables automatic and real-time grading. Teachers can instantly receive results and student performance statistics, including item analysis, difficulty level, and achievement level of critical thinking indicators. This accelerates the process of reflection and decision-making for subsequent learning. This finding supports modern assessment principles, which state that a well-designed evaluation instrument should not only measure the mastery of facts but also students' ability to apply higher-order thinking skills such as analysis and evaluation (Stobaugh, 2018). This instrument is not only suitable for summative assessment but is also ideal for use as an initial diagnostic tool or for formative evaluation during instruction. With questions designed based on critical thinking ability indicators, teachers can identify students' thinking weaknesses more specifically and provide more targeted learning interventions. This aligns with research by Brookhart (2010), who argues that effective formative assessment can be used to monitor student progress in developing each facet of critical thinking, allowing educators to design more targeted instructional strategies to foster these skills.

Instrument Feasibility

An assessment instrument is considered feasible if it meets validity criteria, which can be defined as the degree of accuracy of the test instrument used to measure what it is intended to measure (Ramadhani & Fitri, 2020). Expert judgment is a common method utilized by researchers for the validation of newly developed instruments. (Mohamat et al., 2022). Before empirical trials with students, the instrument first underwent an expert validation phase to ensure its feasibility in terms of material, construct, and language. The quantitative outcomes of this evaluation process, as summarized by the experts, are presented in **Table 6**. The expert validation process for this study involved five raters to assess the instrument's content validity. The term 'indicators' refers to the specific checklist items that each expert rated to evaluate the instrument's quality. These indicators were grouped into three main aspects: Material, Construct, and Language. The amount of indicators is not the same for each aspect because each aspect requires a different level of detail for a thorough evaluation. Specifically, the material aspect consisted of five indicators, the construct aspect had nine indicators, and the language aspect had four indicators. The quantitative results of this expert judgment, presented in **Table 6** as Aiken's V coefficients, show that all instrument components surpassed the minimum validity threshold and were declared valid for use.

Table 6. Expert Validity Test of The Instrument

Aspect	∑ Indicator	Average V	V table	Description
material	5	0,93	0,87	valid
construct	9	0,94	0,87	valid
language	4	0,93	0,87	valid

The content validity of the instrument was determined using the Aiken's V coefficient, which measures the level of agreement among experts. The instrument is considered valid because the validation criteria stipulate that the calculated V value must be greater than the minimum threshold value from the Aiken's V table. For this study, which involved five expert raters, the minimum threshold value from the Aiken table was determined to be 0.87. The analysis of the expert ratings yielded an average validity value of V = 0.93 across all aspects (material, construction, and language). Since the obtained average value of 0.93 clearly exceeds the required minimum of 0.87, it is concluded that the developed instrument has met the standards for content validity and is thus suitable for use (Amrianto et al., 2024).

Table 7. Item Validity Test Results

Item	Small-scale Test			Large-scale Test		
	MNSQ	ZSTD	Pt.MS.Corr	MNSQ	ZSTD	Pt.MS.Corr
1	1,61	0,83	0,17	0,74	-0,37	0,36
2	1,57	-0,44	0,57	1,21	1,38	0,65
3	1,45	1,41	0,56	0,99	-0,01	0,70
4	1,07	0,37	0,33	1,43	1,22	0,38
5	1,46	1,37	0,55	1,11	0,63	0,58
6	0,46	-0,49	0,45	0,52	-1,22	0,45
7	0,34	-0,18	0,27	0,76	-0,33	0,33
8	0,80	-0,36	0,79	0,84	-1,09	0,74
9	0,79	0,19	0,27	1,93	1,85	0,36
10	0,80	-0,36	0,79	0,97	-0,17	0,66
11	0,46	-0,73	0,53	1,10	0,38	0,45
12	1,08	0,34	0,34	0,69	-1,09	0,52
13	1,74	1,24	1,74	1,23	0,94	0,50
14	2,38	1,43	2,38	1,74	1,86	0,39
15	0,69	0,08	0,29	0,77	-0,30	0,34
16	1,07	0,37	0,33	1,38	1,40	0,49
17	0,46	-0,73	0,53	0,63	-0,91	0,44
18	0,78	-0,34	0,53	0,89	-0,40	0,57
19	0,57	-0,70	0,54	0,64	-1,20	0,51
20	0,91	-0,15	0,73	0,97	-0,17	0,66

In the empirical trial, the data obtained were analyzed with the Rasch Model using Winsteps software. By using Winsteps software, item validity and respondent validity can be analyzed separately, thus obtaining more accurate data results (Ridwan et al., 2023). Item validity was analyzed based on three criteria: outfit mean square (MNSQ), outfit Z-standard (ZSTD), and Point Measure Correlation (Pt. Measure Corr) with item fit or misfit (outlier) criteria as follows based on Boone et al (2014): (1) The Outfit MNSQ value should be > 0.5 and < 1.5, and the closer to 1 the better; (2) The Outfit ZSTD value should be > -2.0 and < 2.0, the closer to 0 the better; and (3) the Pt Measure Corr value > 0.4 and < 0.85. An item is considered fit if it meets at least one of the three criteria. If an item is found where the MNSQ and Pt Measure Corr values do not meet the criteria but the ZSTD value does, the item is still considered fit, meaning the item is retained. Following expert validation, the instrument was empirically tested and analyzed using the Rasch Model to assess the fit of each item. **Table 7** provides a detailed breakdown of the item validity analysis, presenting key statistical values, including MNSQ, ZSTD, and Point Measure Correlation, for all 20 items. This data is displayed for both the small-scale and large-scale trials, allowing for a direct comparison of how each item functioned within different sample groups. To ensure a comprehensive and robust validation of the instrument, this study employed a two-stage approach that assessed both content validity and empirical validity. The use of both expert judgment and statistical analysis provides a comprehensive and robust validation of the instrument, a practice supported as essential for developing high-quality educational assessments (Retnawati, 2016).

First, content validity was established using Aiken's V. This method was chosen to quantitatively measure the degree of agreement among a panel of experts regarding the relevance, clarity, and representativeness of the instrument's items. This initial step is crucial to confirm that the instrument is theoretically sound and its content is appropriate before being tested on students. Second, empirical validity was determined using the Rasch Model. This statistical model was applied to analyze the actual response data from student trials, providing evidence of how each item performed in a real-world setting and whether it fit the underlying measurement construct (Mohamat et al., 2022).

The findings from both distinct methods led to the same conclusion: the instrument is valid. The expert-based content validation through Aiken's V confirmed the instrument's feasibility. This result was subsequently supported by the data-driven empirical validation from the Rasch Model, which also concluded that all items were valid based on statistical fit. Therefore, the convergence of these results provides strong, comprehensive evidence that the developed instrument is valid and suitable for use.

Instrument reliability is evident in the summary statistics output. Reliability refers to the consistency of measurement; a reliable instrument remains consistent and stable over time, and the instrument has reliability as a measuring tool. Beyond individual item validity, the overall consistency and stability of the instrument are crucial measures of its quality. **Table 8** summarizes the results of the item reliability test, presenting the Item Reliability and Cronbach's Alpha coefficients for both the small-scale and large-scale tests. The high values in this table collectively affirm that the developed instrument possesses an excellent degree of reliability and can be trusted for consistent measurement.

Table 8. Item reliability test results

Test Scale	Item Reliability	Cronbach Alpha	Conclusion
small-scale	0,74	0,84	reliable
large-scale	0,92	0,87	reliable

In the small-scale test, the item reliability value was 0.74. In the large-scale test, the item reliability was much better than in the small-scale test, with a value of 0.92. This indicates that the consistency of answers from students is considered reliable/consistent. The items shown in the instrument have successfully provided a good picture of the students' condition. The Cronbach's Alpha (KR-20) from the small-scale test was 0.84, while in the large-scale test, it was 0.87. This indicates that the developed instrument has a good reliability coefficient and can be relied upon to measure the construct accurately and consistently (Astuti et al., 2022). This indicates that the developed instrument has a high reliability coefficient. The advantage of the Rasch Model is that it enhances the accuracy of the instrument and the consistency of measurement results, allowing for item calibration and a more valid interpretation of research findings (Sujatmika et al., 2025). The achievement of high reliability values indicates the success of the instrument development process, as guided by the 4D model.

The difficulty level of the items can be identified in the Winsteps program by the logit value in the item measure order output column. A good assessment instrument should feature items with a well-distributed range of difficulty to measure student abilities at various levels accurately. **Table 9** presents the results of the item difficulty analysis, where the logit value and corresponding difficulty category (Easy, Medium, Hard) are displayed for each item. This table illustrates that the developed instrument successfully achieves a balanced distribution of item difficulty levels.

Table 9. Item difficulty level test results

Small-scale Test		Large-scale Test	
Category	**Item**	**Category**	**Item**
very easy	-	very easy	-
easy	1,4,6,7,9,11,12,14,15,18,20	easy	1,6,7,9,13,15,17
medium	2,5,13,16,17	medium	4,11,12,14,16,18,19
hard	3,8,10,19	hard	2,3,5,8,10,20
very hard	-	very hard	-

From the table above, it can be seen that the distribution of difficulty levels in the items is evenly spread, from easy to challenging levels. This result is considered very good because no item is classified as very easy or very hard. Items that are too easy tend not to stimulate problem-solving efforts. Conversely, items that are too difficult can make children feel hopeless and less enthusiastic about trying again (Miftahussa'adiah et al., 2020).

The item discrimination power was determined by examining the separation index in the summary statistics output. A critical quality of an assessment item is its ability to differentiate between students with higher and lower abilities on the measured construct, known as item discrimination. To evaluate this characteristic, an analysis was performed to determine the instrument's item separation value. As per your instruction, **Table 10** summarizes the results of this analysis, presenting the item separation index and its interpretation ('Poor' or 'Good') for both the small-scale and large-scale tests. This data is essential for judging the instrument's effectiveness in distinguishing between different levels of student proficiency.

Table 10. Item Discrimination Test Results

Test	Item Separation	Conclusion
small-scale test	1.68	poor
large-scale test	3.44	good

The item discrimination analysis revealed a significant difference in the instrument's performance between the two trial phases. In the small-scale test, the instrument had an item separation value of 1.68, indicating 'weak' discrimination power. In contrast, the value improved substantially to 3.44 in the large-scale test, which is categorized as 'good' discrimination power.

The reason for this difference is primarily attributed to the sample size. The large-scale test involved a significantly larger number of respondents, which provided a wider and more varied distribution of student abilities. According to the principles of psychometric analysis, a larger sample size allows for a more stable and accurate estimation of item parameters within the Rasch Model. This stability leads to a more reliable calculation of the item separation index, which explains the improvement from the small-scale to the large-scale test. This finding is consistent with research indicating that Rasch Model analysis yields more robust and reliable results with larger samples (Mohamat et al., 2022).

To detect item bias in this study, the Mantel-Haenszel method was used (Mantel & Haenszel, 1959). Two groups were analyzed: male and female. To ensure fairness, the instrument was analyzed for item bias to determine if any questions unfairly favored one group of students over another based on gender. Using the Mantel-Haenszel method, items were flagged as potentially biased if they met specific statistical criteria, including having a probability value of less than 5%. **Table 11** presents the results of this differential item functioning (DIF) analysis, showing the probability values for each of the 20 items across both test scales to detect any significant performance differences between male and female students.

For the bias test on a small scale, the probability values for all items were above 0.05 (i.e., greater than 5%). This leads to the conclusion that the items in the instrument are free from bias. In the large-scale bias test, it was found that only question number 2 (prob. 0.0130) was detected as biased. In contrast to the large-scale test, item Q2 ($p = 0.0130$) was found to contain bias because its p-value was very small, far below the 0.05 threshold. This is worth noting, considering that items other than Q2 still yielded good probability values far above 0.05.

In addition to the technical aspects of the items, the effectiveness of the instrument is also seen from its use in learning practices. Based on interviews with teachers and observations in the classroom, it was found that the instrument effectively encouraged students to think critically and analytically. This finding supports modern assessment principles, which state that a well-designed evaluation instrument should measure not only the mastery of facts but also students' ability to apply higher-order thinking skills such as analysis and evaluation (Stobaugh, 2018).

Ultimately, an instrument's effectiveness is judged by its ability to measure student achievement on the targeted skills. Table 12 presents this crucial student performance data, showing the percentage of correct answers for each of the six critical thinking skill indicators during the small-scale and large-scale tests. The data in this table allows for a direct evaluation of the instrument's practical utility in diagnosing specific student competencies in each aspect of critical thinking. This increase was quite significant in the large-scale test, where the percentage values tended to be higher than those in the small test. However, on the interpretation indicator, the percentage in the small-scale test was slightly higher than in the large-scale test. This difference is not too significant, considering that the other five indicators in the large-scale test had a pretty significant percentage advantage. Overall, the effectiveness of the instrument in this study is relatively high, as viewed from various aspects. This finding aligns with previous research by Fauzi & Wicaksono (2021), which stated that a systematically designed evaluation instrument that has undergone adequate validity and reliability tests will provide objective and accurate assessment results.

Table 11. Item Bias Test Results

Item	Small-scale Test		Large-scale Test	
	Prob.	Conclusion	Prob.	Conclusion
1	0.7650	not biased	0.3171	not biased
2	0.8111	not biased	0.0130	biased
3	0.8765	not biased	0.8427	not biased
4	0.7880	not biased	0.2249	not biased
5	0.6194	not biased	0.5380	not biased
6	0.2070	not biased	0.3548	not biased
7	0.3477	not biased	0.7733	not biased
8	0.9743	not biased	0.3818	not biased
9	0.1478	not biased	0.1645	not biased
10	0.9745	not biased	0.1166	not biased
11	0.6450	not biased	0.1276	not biased
12	0.6450	not biased	0.4618	not biased
13	0,.2875	not biased	0.2906	not biased
14	0.7880	not biased	0.3600	not biased
15	0.3426	not biased	0.3512	not biased
16	0.2070	not biased	0.0993	not biased
17	0.6450	not biased	0.2877	not biased
18	0.6821	not biased	0.7294	not biased
19	1.0000	not biased	0.4783	not biased
20	0.7440	not biased	0.1166	not biased

Table 12. Student performance results

Indicator	% Correct Answer	
	Small-scale Test	Large-scale Test
interpretation	77.50	76.17
analysis	76.67	84.89
evaluation	62.22	73.33
inference	74.67	86.00
explanation	74.44	82.89
self-regulation	65.00	75.67

The data presented in Table 12 indicates that the overall student performance on the critical thinking skills instrument was good, with an average mastery level of 79.82%. This suggests that the instrument is

effective in measuring the intended skills. However, a deeper analysis of the results reveals a notable variation in student proficiency across the different cognitive indicators.

An examination of student answer sheets provides insight into this disparity. The highest level of mastery was observed in the inference indicator (86.00%), while the lowest was in the evaluation indicator (73.33%). For items measuring 'inference,' students were generally successful in applying concepts like Ohm's law to correctly calculate and conclude relationships from the data provided. In contrast, on items measuring 'evaluation,' many students struggled. The answer sheets showed that they had difficulty providing logical and relevant arguments to support their claims, often stating simple agreement or disagreement without the strong scientific justification required for higher-order thinking (Ennis, 2018).

These findings highlight the diagnostic utility of the developed instrument. By identifying specific areas of weakness, such as 'evaluation,' educators can use this test to design targeted learning interventions. For example, a teacher could provide more focused exercises that require students to critique arguments or justify conclusions with evidence, thereby hopefully increasing that skill over time. Furthermore, the instrument is not only useful as a summative tool at the end of a unit but can also be effectively employed as a formative assessment to periodically monitor students' progress in developing each facet of critical thinking throughout the learning process (Brookhart, 2010).

Conclusion

This study aimed to develop a feasible, valid, and effective instrument for assessing the critical thinking skills of high school students in dynamic electricity. Based on the comprehensive validation and testing process, the goal of this research was successfully achieved. The developed instrument demonstrated strong psychometric properties, confirming its suitability for use in an educational setting. The achievement of this goal is supported by several key data points. First, the instrument's content validity was affirmed by five experts, yielding a high average Aiken's V coefficient of 0.93. Second, empirical analysis using the Rasch Model confirmed its quality, showing excellent reliability (Cronbach's Alpha = 0.87 for the large-scale test) and good item discrimination (item separation index = 3.44). The instrument also proved effective in practice, as students achieved an average mastery score of 79.82%, indicating that the items were well-calibrated for the target population. The successful development of this instrument has several practical implications for educators. This tool can be used not only for summative evaluation but also as a diagnostic and formative assessment to identify specific areas where students struggle, such as the 'evaluation' skill. By providing detailed feedback on student performance across different cognitive dimensions, teachers can design more targeted instructional strategies to foster higher-order thinking. For further research, several avenues are suggested. First, future studies could expand the development of similar instruments for other physics topics to create a comprehensive assessment suite. Second, research could be conducted on a more diverse and widespread population to further test the instrument's generalizability. Finally, a longitudinal study could be undertaken to investigate whether the regular use of this instrument as a formative tool directly contributes to an increase in students' critical thinking skills over an academic year.

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142.

Amrianto, Rohman, F., Dharmawan, A., & Sari, M. S. (2024). Development of a 4C skills evaluation instrument for biology: A validity and reliability study on Indonesian high school students learning. *International Journal of Innovative Research and Scientific Studies, 7*(2), 701–717. https://doi.org/10.53894/ijirss.v7i2.2873

Arifin, Z. (2019). *Evaluasi pembelajaran* (P. Latifah, Ed.). ROSDA.

Asriadi, M. A. M., & Hadi, S. (2021). Pengembangan tes diagnostik kemampuan berpikir kritis pada mata pelajaran fisika [Master's thesis, Universitas Negeri Yogyakarta]. Universitas Negeri Yogyakarta Repository.

Astuti, B., Purwanta, E., Ayriza, Y., Bhakti, C. P., Lestari, R., & Herwin, H. (2022). School connectedness instrument's testing with the Rasch model for high school students during the COVID-19 pandemic. *Cypriot Journal of Educational Sciences, 17*(2), 410–421. https://doi.org/10.18844/cjes.v17i2.6828

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Desilva, D., Sakti, I., & Medriati, R. (2020). Pengembangan instrumen penilaian hasil belajar fisika berorientasi HOTS (higher order thinking skills) pada materi elastisitas dan hukum Hooke. *Jurnal Kumparan Fisika, 3*(1), 41–50. https://doi.org/10.33369/jkf.3.1.41-50

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi, 37*(1), 165–184. https://doi.org/10.1007/s11245-016-9401-4

Hasan, S. W., Auliah, A., & Herawati, N. (2020). Pengembangan instrumen penilaian kemampuan berpikir kritis siswa SMA. *Chemistry Education Review, 3*(2), 185–195. https://doi.org/10.26858/cer.v3i2.13769

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2024). *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia Nomor 12 Tahun 2024 tentang Kurikulum pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah*.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Miftahussa'adiah, M., Alberida, H., & Handayani, D. (2020). Pengembangan asesmen kemampuan berpikir kritis materi sistem sirkulasi untuk siswa SMA kelas XI. *SIMBIOSA, 9*(1), 39–46. https://doi.org/10.33373/sim-bio.v9i1.2434

Mohamat, R., Sumintono, B., & Abd Hamid, H. S. (2022). Analisis kesahan kandungan instrumen kompetensi guru untuk melaksanakan pentaksiran bilik darjah menggunakan model Rasch pelbagai faset (Content validity analysis of an instrument to measure teacher's competency for classroom assessment using many facet Rasch model). *Jurnal Pendidikan Malaysia, 47*(1), 1–13. https://doi.org/10.17576/JPEN-2022-47.01-01

Ramadhani, R., & Fitri, Y. (2020). Validitas e-modul matematika berbasis EPUB3 menggunakan analisis Rasch model. *Jurnal Gantang, 5*(2), 95–111. https://doi.org/10.31629/jg.v5i2.2535

Retnawati, H. (2016). *Validitas, reliabilitas, dan karakteristik butir*. Parama Publishing.

Ridwan, M. R., Hadi, S., & Jailani, J. (2023). Measurement of psychometric properties numerical aptitude assessment scale for prospective high school students: A Rasch model analysis. *TEM Journal, 12*(4), 2416–2429. https://doi.org/10.18421/TEM124-54

Setiawan, R. (2024). Optimasi pengalaman pengguna: Penilaian otomatis dan pencegahan kecurangan ujian online. *Bit-Tech, 7*(2), 299–306. https://doi.org/10.32877/bt.v7i2.1758

Sihombing, B., Zamsiswaya, & Sawaluddin. (2024). Model Pengembangan 4D (Define, Design, Develop, dan Disseminate) dalam Pembelajaran Pendidikan Islam. *Journal of Islamic Education El Madani, 4*(1), 11–19. https://doi.org/10.55438/jiee.v4i1.135

Sivakumar, S., Vidyanandini, S., Nayak, S. R., & Aluvala, S. (2024). Modular irregular labelling in Network Analysis with Complete Bipartite Graph as Auto Proctoring. *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, 589–592. https://doi.org/10.1109/ESIC60604.2024.10481612

Standar Kompetensi Lulusan, Pub. L. No. Permendikbud No 20. Tahun 2016, Kementrian Pendidikan dan Kebudayaan 1 (2016).

Stobaugh, R. (2018). *Assessing critical thinking in middle and high schools: A standards-based approach*. Routledge.

Sugiyono. (2017). *Metode Penelitian Kuantitatif Kualitatif R&D*. Alfabeta.

Sujatmika, S., Sutarno, S., Masykuri, M., & Prayitno, B. A. (2025). Applying the Rasch model to measure students' critical thinking skills on the science topic of the human circulatory system. *Eurasia Journal*

of Mathematics, Science and Technology Education, 21(4), em2622. https://doi.org/10.29333/ejmste/16221

Waruwu, M. (2024). Metode Penelitian dan Pengembangan (R&D): Konsep, Jenis, Tahapan dan Kelebihan. *Jurnal Ilmiah Profesi Pendidikan*, 9(2), 1220–1230. https://doi.org/10.29303/jipp.v9i2.2141