Journal of Educational Sciences
Journal homepage: https://jes.ejournal.unri.ac.id/index.php/JES
P-ISSN 2581-1657
E-ISSN 2581-2203

Development of Critical Thinking and Logical Argumentation Assessment Instruments for Students in Problem-Based Learning

Muhammad Khoirul Fikri*, Rusijono, Khusnul Khotimah
Department of Educational Technology, Universitas Negeri Surabaya, Surabaya 60213, Indonesia

ARTICLE INFO

Article history:

Received: 25 Nov 2025
Revised: 19 Des 2025
Accepted: 24 Des 2025
Published online: 05 Jan 2026

Keywords:

Instrument development; critical thinking; logical argumentation; Problem-Based Learning

* *Corresponding author:*

E-mail: 24010905010@unesa.ac.id

Article Doi:

https://doi.org/10.31258/jes.10.1.p.1391-1404

This is an open access article under the CC BY-SA license.

A B S T R A C T

The background of this study is the demand of the Industrial Revolution 4.0 for Higher Order Thinking Skills (HOTS), particularly critical thinking and logical argumentation among university students. Problem-Based Learning (PBL) is recognized as effective in developing these skills; however, valid and comprehensive assessment instruments, especially in essay form, remain limited. This study aims to develop and examine the quality of an essay test instrument to measure students' critical thinking and logical argumentation skills within PBL in the Logic and Critical Thinking Methods course. The method employed was Research and Development (R&D), adapting a simplified Borg and Gall model, which included needs analysis, test blueprint design, item development, expert validation, empirical try-outs, and item quality analysis. The research subjects were 55 students in Surabaya. Expert validation using Aiken's V showed an average coefficient of 0.81, categorized as very high. Construct validity testing using item-total correlation confirmed that all items were valid, and reliability testing using Cronbach's Alpha yielded a coefficient of 0.768, indicating high reliability. Item quality analysis revealed that most items possessed moderate difficulty and high discrimination power. In conclusion, the developed instrument is valid, reliable, and suitable for use as a measurement tool for assessing students' critical thinking and logical argumentation abilities in a PBL context.

1. Introduction

The advent of the Industrial Revolution 4.0 and Society 5.0 necessitates that university students possess not only factual knowledge but, more critically, Higher Order Thinking Skills (HOTS). Among the core competencies that must be cultivated in higher education is critical thinking ability, defined as the skill to

analyze information, evaluate arguments, and make logical, evidence-based decisions (Arwaty et al., 2023; Nuryanti et al., 2018). This competency is even positioned as a learning outcome indicator within the Indonesian National Qualification Framework (KKNI), particularly in general courses such as Logic and Critical Thinking Methods, which directly aim to shape students' analytical and systematic thinking patterns. In an increasingly complex world saturated with misinformation, critical thinking skills serve as a vital shield for students to assess the credibility of information sources and digital claims amidst a torrent of often unverified information (Sulastri & Cahyani, 2021).

According to Facione (2015), critical thinking comprises six key indicators: interpretation, analysis, evaluation, inference, explanation, and self-regulation. These indicators play a central role in the information processing process, enabling individuals not merely to receive information but also to deconstruct, weigh, conclude, and reconstruct it into coherent and well-founded arguments (Reynders et al., 2020). This process is the core of logical argumentation ability, where students are required to present their thoughts in a structured manner, supported by relevant reasons and strong evidence. Therefore, the measurement of critical thinking must also encompass the aspect of argument construction, as the two are two sides of the same coin in quality reasoning (McGrew et al., 2018).

In fostering these skills, Problem-Based Learning (PBL) is recognized as an effective instructional model. PBL places students in collaborative, authentic problem-solving situations that require them to conduct inquiry, analyze data, and construct arguments to propose solutions. This inquiry cycle in PBL is inherently aligned with Facione's critical thinking indicators, where students actively interpret problems, analyze their components, evaluate available information, make inferences, explain their reasoning, and engage in self-regulation of their own thinking processes (Hmelo-Silver, 2004; Savery, 2006). Thus, the implementation of PBL in the Logic and Critical Thinking Methods course is not only relevant but strategic for honing students' reasoning competencies in an integrated and contextual manner.

However, the effectiveness of PBL in enhancing critical thinking and logical argumentation skills requires valid and reliable assessment instruments to measure it. Unfortunately, previous studies indicate that instruments for measuring critical thinking, particularly in the context of such courses, remain limited and not comprehensive (Tiruneh & Cock, 2017). Many existing instruments still rely on multiple-choice formats, which tend to restrict the measurement of higher-order thinking dimensions. Objective tests often only cover aspects of memory or basic comprehension and are less capable of revealing students' capacity for in-depth analysis, complex evaluation, let alone constructing well-structured and logical arguments in writing (Stanger-Hall, 2012). Consequently, the resulting portrayal of students' critical thinking abilities is incomplete.

Based on this gap, the development of essay-based test instruments becomes imperative. Essay instruments provide space for students to articulate their thinking processes, logical flow, and how they construct justifications in depth, making them

more suitable for measuring complex constructs like critical thinking and argumentation. Such an instrument needs to be designed based on comprehensive indicators, validated by experts, and empirically tested to ensure its quality as a measurement tool. Therefore, this study aims to develop and examine the quality of an essay test instrument designed to measure students' critical thinking skills and logical argumentation abilities within Problem-Based Learning. The developed instrument is based on Facione's critical thinking framework and incorporates indicators of logical argument construction, with the goal of producing a valid, reliable, and comprehensive assessment tool for evaluating higher-order thinking outcomes in PBL-based learning contexts in higher education.

2. Methodology

This study employed a Research and Development (R&D) approach by adapting a simplified Borg and Gall development model, integrated with the test instrument development framework proposed by Ramadhan et al. (2019) to ensure instrument quality from indicator formulation, expert validation, and try-out stages to item analysis. The developed test instrument was oriented toward assessment development to measure students' critical thinking skills within the context of Problem-Based Learning (PBL).

The development process was conducted through several stages, including needs analysis, indicator formulation, item construction, expert validation, instrument try-out with students, and item quality analysis. This approach was selected to produce a test instrument that is not only theoretically sound but also empirically proven to be valid and reliable for evaluating student learning outcomes. The research subjects consisted of 55 students enrolled in the Logic and Critical Thinking Methods course at a higher education institution in Surabaya. The subjects were selected using purposive sampling, based on the consideration that these students had acquired fundamental reasoning and argumentation skills necessary to respond to case-based questions requiring critical analysis.

The research instrument was an essay test consisting of 20 items, developed based on six critical thinking indicators proposed by Facione (2015) interpretation, analysis, evaluation, inference, explanation, and self-regulation and five indicators of logical argument construction: clarity of claims, relevance of reasons, strength of evidence, logical argument structure, and the ability to address counterarguments. All items were designed using authentic problem contexts relevant to students' real-life experiences to ensure the assessment measured higher-order thinking processes rather than mere factual recall.

Research data were collected through the administration of a test at the end of the PBL-based learning activities. The test implementation was directly supervised by the researcher to prevent procedural bias and collaboration among students during test completion. In addition, content validation was conducted by two experts—one subject-matter expert in logic and one expert in educational evaluation—to ensure the appropriateness of content, item construction, and language clarity of the

instrument. The results of expert validation were analyzed using Aiken's V coefficient, with an instrument considered valid when the coefficient value reached a minimum of 0.80. The formula used to calculate content validity using Aiken's V is as follows:

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

V: Aiken's V coefficient
S: The adjusted expert score was obtained by subtracting the lowest score on the rating scale, expressed as $s = r - l_0$, where $r$represents the score assigned by the expert and $l_0$denotes the lowest possible score on the scale.
r: score given by the expert
c: number of rating scale categories
n: number of experts

Table 1. Criteria for Aiken's V

Aiken's V Value	Category
V≤ 0,4	Low
0,4 < V ≤ 0,8	Moderate
0,8 < V ≤ 1,0	High

(Utami et al., 2024)

After expert validation, the instrument was administered in a try-out to obtain empirical data, which were subsequently analyzed to determine the quality of the test items. Item validity analysis was conducted using item–total correlation through the Pearson Product Moment correlation. An item was considered empirically valid if the calculated correlation coefficient ($r_{calculated}$) exceeded the critical value (r_{table}) at a significance level of 0.05.
The formula used to calculate the Pearson Product Moment correlation is as follows:

$$rxy = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

Construct validity analysis was conducted using SPSS with the following criteria:

Table 2. Criteria for Construct Validity

Construct Validity Value (D)	Category
0,7 < D ≤ 1,0	Very High
0,4 < D ≤ 0,7	High
0,2 < D ≤ 0,4	Moderate
0 < D ≤ 0,2	Low

(Sholeha et al., 2023)

Furthermore, the reliability of the instrument was calculated using Cronbach's Alpha, with a minimum reliability coefficient threshold of 0.70, indicating adequate

internal consistency in measuring the same construct. The criteria for the reliability test are presented in Table 3 below.

Table 3. Instrument Reliability Criteria

Reliability Coefficient	Category
0,81-1,00	Very High
0,61-0,80	High
0,41-0,60	Moderate
0,21-0,40	Low
0,00-0,20	Very Low

(Nu'man, 2023)

All statistical analyses were conducted using the latest version of SPSS software to ensure the accuracy of the calculations. Through these methodological procedures, the developed instrument is expected to serve as a high-quality evaluation tool for comprehensively measuring students' critical thinking achievement within the context of Problem-Based Learning, particularly in the Logic and Critical Thinking Methods course.

3. Result And Discussion

Needs Analysis and Preliminary Study Findings

The development of this instrument was initiated through a comprehensive needs analysis and preliminary study, which identified a significant gap in the assessment tools available for measuring higher-order thinking skills within student-centered learning environments. The analysis involved a review of current literature, an evaluation of existing assessment practices in the Logic and Critical Thinking Methods course, and consultations with course instructors. The findings consistently revealed that while Problem-Based Learning (PBL) was recognized as an effective pedagogical model for fostering critical thinking and argumentation, its assessment primarily relied on conventional methods. These methods, often multiple-choice or short-answer formats, were found to be insufficient for capturing the complex, multi-faceted, and process-oriented nature of critical thinking and logical argument construction.

The preliminary study, based on a synthesis of theoretical frameworks, established a firm foundation for instrument development. It confirmed the relevance and comprehensiveness of Facione's (2015) six core critical thinking indicators—interpretation, analysis, evaluation, inference, explanation, and self-regulation—as the primary cognitive dimensions to be measured. Furthermore, the analysis underscored that the ability to think critically is intrinsically linked to the capacity for sound argumentation. Therefore, the study integrated five key indicators of logical argumentation—clarity of claim, relevance of reasons, strength of evidence, logical structure, and the ability to address counterarguments—into the construct to be assessed. This integrated framework ensured the instrument would measure not just discrete cognitive skills but also the holistic ability to produce reasoned and

structured written arguments. The synthesis of these findings made a compelling case for the necessity of developing a dedicated, essay-based test instrument specifically tailored to the contextual and cognitive demands of PBL, thereby addressing the identified gap in valid and reliable assessment tools for these essential 21st-century skills.

Planning and Development of the Preliminary Product

This study successfully developed an essay-based test instrument designed to comprehensively measure students' critical thinking skills and logical argumentation abilities within a Problem-Based Learning (PBL) context for the Logic and Critical Thinking Methods course. The final instrument consists of 13 essay items, each constructed to elicit and assess higher-order cognitive processes beyond mere factual recall. The instrument was developed based on a detailed test blueprint that meticulously aligned each item with the core constructs being measured. Specifically, the items were designed to correspond with six critical thinking indicators according to Facione (2015), interpretation, analysis, evaluation, inference, explanation, and self-regulation, and were contextualized within five key aspects of logical argumentation: clarity of claims, relevance of reasons, strength of evidence, logical structure of arguments, and the ability to address counterarguments.

Each essay item presents an authentic, case-based problem scenario relevant to students' academic and real-world experiences, requiring them to analyze information, evaluate different perspectives, draw reasoned conclusions, and articulate their reasoning in a structured written form. To ensure objective and consistent scoring, a detailed analytic rubric on a scale of 1 to 5 was developed for each item. This rubric explicitly defines performance levels for each assessed criterion (e.g., depth of analysis, soundness of evaluation, coherence of argument structure), thereby reducing subjective bias in evaluation. The following blueprint (Table 1) summarizes the mapping of each test item to the primary critical thinking indicator and argumentation aspect it targets, illustrating the instrument's comprehensive coverage of the intended constructs. The following is the Test instrument blueprint mapping items to critical thinking and argumentation constructs in table 4:

Table 4. Test Instrument Blueprint Mapping Items to Critical Thinking and Argumentation Constructs

Item Number	Problem Scenario Context	Primary Critical Thinking Indicator	Key Argumentation Aspect Assessed
1.	Analyzing a controversial editorial	Interpretation	Clarity of Claim
2.	Evaluating scientific vs. anecdotal evidence in a public health issue	Evaluation	Strength of Evidence
3.	Solving a logic puzzle with incomplete data	Inference	Logical Structure
4.	Debating ethical implications of a technology	Analysis	Relevance of Reasons

5.	Reflecting on a personal reasoning bias in a case study	Self-regulation	Addressing Counterarguments
6.	Explaining a complex social phenomenon	Explanation	Clarity of Claim & Logical Structure
7.	Comparing two conflicting research summaries	Analysis & Evaluation	Strength of Evidence & Relevance of Reasons
8.	Proposing a solution to a campus resource dilemma	Inference & Explanation	Logical Structure & Addressing Counterarguments
9.	Critiquing a flawed statistical claim in media	Evaluation	Strength of Evidence
10.	Interpreting data trends from a graph	Interpretation & Analysis	Clarity of Claim
11.	Constructing an argument for a policy change	Inference & Explanation	Relevance of Reasons & Logical Structure
12.	Self-assessing the reasoning process in a previous argument	Self-regulation	Addressing Counterarguments
13.	Synthesizing multiple sources on a historical event	All indicators (Holistic)	All aspects (Comprehensive)

Preliminary Field Testing (Expert Validation)

The results of the content validity analysis indicated that the Aiken's V coefficients for individual items ranged from 0.58 to 1.00, with an average value of 0.81, which falls into the very high category. This finding indicates a high level of agreement among validators regarding the alignment of the instrument content with the indicators of critical thinking and logical argumentation skills. The results of the content validity analysis of the essay test instrument (Aiken's V) from the three validators are presented in the following table 5.

Table 5. Results of Content Validity

Item	**Validator I**	**Validator II**	**Validator III**	**Σs**	**n(c−1)**	**V**	**Category**
Item 1	4	5	4	10	12	0,83	Very High
Item 2	3	5	4	9	12	0,75	High
Item 3	4	4	4	9	12	0,75	High
Item 4	5	4	4	10	12	0,83	Very High
Item 5	5	4	5	11	12	0,92	Very High
Item 6	4	4	5	10	12	0,83	Very High
Item 7	4	3	4	8	12	0,67	High
Item 8	3	4	4	7	12	0,58	Moderately High
Item 9	5	5	5	12	12	1,00	Very High
Item 10	4	4	4	9	12	0,75	High
Item 11	5	4	4	10	12	0,83	Very High
Item 12	4	3	4	8	12	0,67	High
Item 13	5	4	5	11	12	0,92	Very High
Rata-rata						**0,81**	**Very High**

Thus, based on the validators' assessments and data analysis using Aiken's V coefficient, it can be concluded that the developed essay test instrument meets the

criteria for high content validity. The average Aiken's V value of 0.81 indicates a strong level of agreement among validators regarding the alignment of the test items with the indicators of critical thinking skills and logical argumentation ability, the clarity of language, and the appropriateness of the scoring guidelines. In the subsequent stage, the instrument that had been declared content-valid was administered to students for further analysis through construct validity and reliability testing using SPSS software. Construct validity testing was conducted using the Pearson Product Moment correlation by correlating the score of each item with the total test score. The results of the construct validity analysis assisted by SPSS are presented in the following table 6.

Table 6. Construct Validity

Item	r (Item–Total)	Sig.	Category	Decision
P01	0.496	0.000	High	Valid
P02	0.300	0.026	Moderate	Valid
P03	0.455	0.000	High	Valid
P04	0.316	0.019	Moderate	Valid
P05	0.280	0.039	Moderate	Valid
P06	0.396	0.003	Moderate	Valid
P07	0.471	0.000	High	Valid
P08	0.280	0.038	Moderate	Valid
P09	0.339	0.011	Moderate	Valid
P10	0.392	0.003	Moderate	Valid
P11	0.303	0.024	Moderate	Valid
P12	0.516	0.000	High	Valid
P13	0.475	0.000	High	Valid

Based on the analysis results, all instrument items (13 items) demonstrated positive correlation coefficients with significance values of Sig. < 0.05; therefore, all items were declared empirically valid. The item correlation coefficients ranged from 0.280 to 0.516, which fall within the moderate to high categories. Items classified as high indicate a strong association with the measured construct, while items in the moderate category still reflect an adequate contribution to measuring students' critical thinking skills and logical argumentation abilities. The variation in correlation values across items suggests that the instrument is not redundant but instead measures multiple aspects of higher-order thinking skills comprehensively. These findings indicate that each item is capable of representing the intended construct and is aligned with the objectives of instrument development within the context of Problem-Based Learning.

Main Product Revision (Based on Expert Feedback)

Following the expert validation process, the preliminary instrument underwent a systematic revision phase to refine its content, structure, and usability. The feedback from the three validators was analyzed thematically, focusing on suggestions for improving item clarity, rubric precision, and contextual relevance. The main revisions implemented were as follows:

1. Linguistic and Contextual Refinement: Based on expert input regarding language clarity, several item stems were rephrased to eliminate ambiguity and ensure all students interpreted the questions as intended. For instance, the context of problem scenarios in Items 2 (public health evidence) and 8 (campus resource dilemma) were made more specific and grounded in relatable academic situations to enhance authenticity and engagement.
2. Rubric Specification: The validators suggested enhancing the discriminatory power of the scoring rubric by making the descriptors for each performance level (scores 1, 3, and 5) more behaviorally anchored. Revisions were made to provide clearer, observable benchmarks for criteria such as "strength of evidence" and "addressing counterarguments," thereby increasing scoring objectivity and consistency.
3. Structural Adjustment: One validator noted a potential overlap in the cognitive demand between two items. Consequently, the focus of Item 6 was sharpened from general explanation to emphasizing the *logical structuring* of an explanation, thereby better distinguishing it from other items focusing on inference or analysis.

No items were removed as a result of this revision, as the content validity indices were predominantly in the high to very high categories. The revision process was thus focused on clarification and specification rather than major overhauls. The outcome was an improved version of the instrument that maintained its theoretical foundation and comprehensive coverage while incorporating expert insights to bolster its validity and practical applicability for the subsequent main field testing. This revised instrument was the version administered to the student sample for empirical validation.

Main Field Testing (Empirical Try-Out)

Instrument reliability was analyzed using Cronbach's Alpha coefficient. The results showed a Cronbach's Alpha value of 0.768, which falls within the reliable category, as described in table 7.

Table 7. Instrument Reliability Results

Coefficient	Value	Number of items	Category
Cronbach's Alpha	0.768	13	Reliable (High)

The reliability value indicates that the instrument has good internal consistency and is capable of providing stable measurement results. According to (DeVellis, 2017), a Cronbach's Alpha value greater than 0.70 indicates acceptable reliability, particularly for instruments measuring complex skills such as critical thinking and logical argumentation. Therefore, the developed instrument is not only construct-valid but also reliable for use in research and learning evaluation.

Operational Product Revision (Item Quality Analysis)

To further evaluate the quality of the 13 essay items beyond validity and reliability, an analysis of item difficulty and discrimination was conducted. This analysis

provides crucial information on how challenging each item was for the students and how effectively each item could differentiate between students with high and low overall critical thinking and argumentation abilities.

The difficulty index for each essay item was calculated based on the ratio of the students' average score on an item to the maximum possible score for that item. Given the use of an analytic rubric (scale 1-5), the formula used was:
P = (Mean Score per Item) / (Maximum Possible Score)
A higher P-value indicates an easier item. The interpretation criteria adapted for this essay instrument are presented in table 8.

Table 8. Difficulty Index (P) Criteria for Essay Items

Difficulty Index (P) Range	Category
$P < 0.30$	Difficult
$0.30 \leq P \leq 0.70$	Moderate
$P > 0.70$	Easy

The calculated difficulty indices for all items are summarized in Table 8. The results show that the majority of items (9 out of 13) fell within the moderate difficulty range ($0.30 \leq P \leq 0.70$), with P-values between 0.42 and 0.65. This indicates that most questions were appropriately challenging for the student sample. Two items (Item 5 and Item 8) were categorized as difficult ($P < 0.30$), with indices of 0.28 and 0.25, respectively. These items involved complex tasks of self-regulation and addressing counterarguments. Conversely, two items (Item 1 and Item 10) were categorized as easy ($P > 0.70$), with indices of 0.78 and 0.74, likely because they focused on foundational interpretation skills. The following are the results of the item difficulty and item discrimination analysis, as shown in table 9:

Table 9. Results of Item Difficulty and Discrimination Analysis

Item Number	Mean Score (Max=5)	Difficulty Index (P)	Difficulty Category	Discrimination Index (D)	Discrimination Category
1.	3.90	0.78	Easy		Moderate
2.	2.65	0.53	Moderate		High
3.	2.10	0.42	Moderate		High
4.	2.95	0.59	Moderate		High
5.	1.40	0.28	Difficult		Moderate
6.	3.25	0.65	Moderate		High
7.	2.35	0.47	Moderate		High
8.	1.25	0.25	Difficult		Moderate
9.	2.80	0.56	Moderate		High
10.	3.70	0.74	Easy		Moderate
11.	2.50	0.50	Moderate		High
12.	2.15	0.43	Moderate		High
13.	2.05	0.41	Moderate		High

The discrimination index (D) measures an item's ability to differentiate between students in the upper group (top 27% of total scores) and the lower group (bottom 27% of total scores). It was calculated using the formula:

D = (Mean Score Upper Group - Mean Score Lower Group) / (Maximum Possible Score).
The interpretation criteria for the discrimination index are shown in table 9. The rules for discrimination analysis (D) in this study are determined according to table 10 below:

Table 10. Discrimination Index (D) Criteria

Discrimination Index (D) Range	Category
D < 0.40	High (Very Good Discriminator)
0.30 ≤ D < 0.40	Moderate (Adequate Discriminator)
D < 0.3	Low (Poor Discriminator)

As presented in Table 8, the analysis yielded positive results for the instrument's discriminative power. Nine items (Items 2, 3, 4, 6, 7, 9, 11, 12, 13) demonstrated high discrimination (D ≥ 0.40), with indices ranging from 0.42 to 0.58. This indicates that these items are excellent at distinguishing between high-ability and low-ability students. The remaining four items (1, 5, 8, 10) showed moderate discrimination (0.30 ≤ D < 0.40), meaning they still adequately differentiate between groups but with less power. Importantly, no items exhibited low discrimination (D < 0.30), confirming that all items contribute meaningfully to the test's overall ability to measure varying levels of student competency.

The combined analysis of difficulty and discrimination provides a comprehensive picture of item performance. Items with moderate difficulty and high discrimination (e.g., Items 2, 3, 4, 7, 9, 11, 12, 13) are considered of excellent quality, as they are optimally challenging and effectively separate students of different ability levels. Items with extreme difficulty (easy or difficult) and moderate discrimination (Items 1, 5, 8, 10) are still acceptable for inclusion but may be candidates for review in future revisions. For instance, the easy items (1 & 10) could be made more complex, while the difficult items (5 & 8) might require clearer prompts or scaffolding. Overall, the analysis confirms that the 13-item instrument is composed of high-quality questions that are appropriately challenging and possess strong discriminative power for measuring critical thinking and logical argumentation in a PBL context.

Final Product Description

The systematic Research and Development process culminates in the final product: the Critical Thinking and Logical Argumentation Test for Problem-Based Learning (CTLAT-PBL). This operational instrument is a valid, reliable, and practically applicable assessment tool designed for implementation in higher education, specifically within courses such as Logic and Critical Thinking Methods that utilize Problem-Based Learning pedagogy. The instrument successfully integrates theoretical constructs into a measurable format, consisting of thirteen contextual essay items, each accompanied by a detailed analytic scoring rubric and clear administration guidelines.

The final test instrument presents a cohesive series of authentic problem scenarios that require students to apply critical thinking and argumentation skills in an integrated manner. These scenarios cover diverse themes including media analysis, ethical deliberation, resource allocation, and scientific critique, ensuring broad relevance and engagement. Each item is carefully constructed to elicit higher-order cognitive processes aligned with Facione's critical thinking indicators and key aspects of logical argumentation, moving beyond simple recall to assess analysis, evaluation, inference, and structured reasoning.

A comprehensive, criterion-referenced analytic rubric supports the scoring process, organized around five core dimensions: clarity of claim or thesis, depth of analysis and interpretation, quality of evaluation and evidence, logical structure and coherence, and self-regulation with attention to counterarguments. Each dimension includes explicit descriptors for performance levels on a scale from one to five, enabling consistent, objective, and transparent evaluation across different raters. Brief administration guidelines are also provided, recommending a timed individual assessment at the conclusion of a PBL cycle, along with practical advice for raters to ensure fidelity in applying the rubric.

Psychometrically, the CTLAT-PBL demonstrates strong properties, including high content validity as reflected by an Aiken's V coefficient of 0.81, satisfactory to high construct validity across all items, good internal consistency reliability with a Cronbach's alpha of 0.768, and favorable item characteristics—most items show moderate difficulty and good discrimination power. As a finalized tool, it is suitable for educators to assess student learning outcomes in PBL settings, for researchers to evaluate instructional interventions, and for further validation studies across varied academic disciplines and learner populations.

4. Conclusion

This study successfully developed and validated an essay-based test instrument specifically designed to measure critical thinking and logical argumentation skills within a Problem-Based Learning context for the Logic and Critical Thinking Methods course. The final instrument, comprising contextual essay items, was constructed based on a well-established theoretical framework integrating core indicators of critical thinking and logical argumentation. Through a systematic research and development process, the instrument demonstrated strong content validity as evidenced by a high level of expert consensus regarding its relevance, clarity, and alignment with the intended constructs. Empirical testing further confirmed its robust construct validity, with all items showing significant correlations to the overall measured skill set, ranging from moderate to strong associations. The instrument also exhibited high internal consistency, proving its reliability as a stable measurement tool. Additional analysis of item characteristics indicated that the majority of questions were optimally challenging and effectively discriminated between varying levels of student ability. Therefore, this instrument stands as a valid, reliable, and practical assessment tool. It is recommended for educators and researchers aiming to evaluate higher-order thinking outcomes in

PBL environments. Future studies should test the instrument with larger and more diverse samples across different disciplines to further establish its generalizability and robustness.

References

Arwaty, Rahmah, A., & Mustamin, K. (2023). Effectiveness of Online Blended Learning and Critical Thinking Skills of Biology Students in Indonesia : Meta- Analysis Study. *Journal of Research in Science Education*, *9*(12), 1284–1291. https://doi.org/10.29303/jppipa.v9i12.5761

Facione, P. A. (2015). Critical Thinking: What It Is and Why It Counts. *Insight Assessment*, *1*(1), 1–30.

Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and How Do Students Learn? *Educational Psychology Review*, *16*(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can Students Evaluate Online Sources? Learning from Assessments of Civic Online Reasoning. *Theory & Research in Social Education*, *46*(2), 165–193. https://doi.org/https://doi.org/10.1080/00933104.2017.1416320

Nu'man, M. (2023). Analisis Struktur Kovarians Indikator Kesehatan pada Lansia yang Tinggal di Rumah Berfokus pada Perasaan Kesehatan Subjektif. *Aleph*, *1*(2), 149–200. https://doi.org/123456789/167638/341506

Nuryanti, L., Zubaidah, S., & Diantoro, M. (2018). Analisis Kemampuan Berpikir Kritis Siswa SMP. *Jurnal Pendidikan: Teori, Penelitian, Dan Pengembangan*, *3*(2), 155–158. https://doi.org/DOAJ-SHERPA/RoMEO-Google Scholar-IPI

Ramadhan, S., Mardali, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The Development of an Instrument to Measure the Higher Order Thinking Skill in Physics. *European Journal of Educational Research*, *8*(3), 743–751. https://doi.org/10.12973/eu-jer.8.3.743

Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to Assess Critical Thinking and Information Processing in Undergraduate STEM Courses. *International Journal of STEM Education*, *7*(9). https://link.springer.com/article/10.1186/s40594-020-00208-5

Savery, J. R. (2006). Overview of Problem-Based Learning: Definitions and Distinctions. *Interdisciplinary Journal of Problem-Based Learning*, *1*(1), 9–20. https://doi.org/https://doi.org/10.7771/1541-5015.1002

Sholeha, D. F., Kusasi, M., & Bakti, I. (2023). Pengembangan Instrumen Soal Higher-Order Thinking Skills (HOTS) Berbasis Google Form untuk Mengukur Hasil Belajar Peserta Didik pada Materi Hidrolisis Garam. *JCAE (Journal of Chemistry And Education)*, *7*(1), 32–48.

Stanger-Hall, K. F. (2012). *Multiple-Choice Exams : An Obstacle for Higher-Level Thinking in Introductory Science Classes*. *11*, 294–306. https://doi.org/10.1187/cbe.11-11-0100

Sulastri, & Cahyani, G. P. (2021). Pengaruh Project Based Learning dengan Pendekatan STEAM Terhadap Kemampuan Berpikir Kritis pada Pembelajaran Online di SMK Negeri 12 Malang. *Jurnal Pendidikan*

Akuntansi (JPAK), *9*(3), 372–379. https://doi.org/https://doi.org/10.26740/jpak.v9n3.p372-379

Tiruneh, D. T., & Cock, M. De. (2017). Measuring Critical Thinking in Physics : Development and Validation of a Critical Thinking Test in Electricity and Magnetism. *International Journal of Science and Mathematics Education*, *39*(5), 663–682. https://doi.org/10.1007/s10763-016-9723-0

Utami, L., Ilahi, D. P., Ratih, A., Yenti, E., Aiken, I., & Isi, V. (2024). Analisis Indeks Aiken untuk Mengetahui Validitas Isi Instrumen Scientific Habits of Mind. *Journal of Research and Education Chemistry (JREC)*, *6*(1), 59–67. https://doi.org/10.25299/jrec.2024.vol6(1).17430

How to cite this article:

Fikri, M. K., Rusijono., & Khotimah, K. (2026). Development of Critical Thinking and Logical Argumentation Assessment Instruments for Students in Problem-Based Learning. *Journal of Educational Sciences*, *10(1)*, 1391-1404.