

VALIDITY AND RELIABILITY OF CRITICAL THINKING SKILLS INSTRUMENT AND DEEP LEARNING GRADUATES PROFILE IN SCIENCE LEARNING

Masringgit Marwiyah Nst^{*1}, Festiyed², Yerimadesi³, Skunda Diliarosta⁴, Febri Yanto⁵, Putri Zuhra⁶

^{1,6}Doctoral Program of Science Education, Padang State University, Padang, Indonesia

^{2,3,4,5}Department of Science Education, Universitas Negeri Padang, Padang, Indonesia

¹Politeknik Adiguna Maritim Indonesia (Poltek AMI) Medan, Medan, Indonesia

*Corresponding author *e-mail: masringgitmarwiyahnst@gmail.com*

Article History

Received: 11 June 2026

Revised: 16 September 2026

Published: 27 September 2026

ABSTRACT

This study aims to test the validity and reliability of two main instruments in the research on the development of a video-assisted PjBL model in science learning in junior high schools. The first instrument is a critical thinking skills test in the form of an essay based on Ennis's taxonomy, and the second instrument is a deep learning graduate profile questionnaire that measures three-character dimensions: independence, collaboration, and creativity (27 statement items). Content validity was tested using the Content Validity Ratio (CVR) and Content Validity Index (CVI) involving 5 expert validators. The reliability of the Critical Thinking Skills instrument was tested with inter-rater reliability (Cohen's Kappa), while the graduate profile questionnaire was tested with Cronbach's Alpha on 30 respondents. The results showed: the KBK instrument obtained a CVI = 1.000 (very valid) with a Kappa coefficient $\kappa = 0.826$ (very reliable); the graduate profile instrument obtained a CVI = 0.852 (valid) with a Cronbach's Alpha $\alpha = 0.946$ (very reliable). All items were declared suitable for use in research, with several items requiring minor revisions based on validator notes.

Keywords: validity; reliability; critical thinking skills; graduate profile; deep learning; Pjbl

Copyright © 2026, The Author(s).

How to cite: Marwiyah Nst, M., Festiyed, Yerimadesi, Diliarosa, S., Yanto, F., Zuhra, P. (2026). VALIDITY AND RELIABILITY OF CRITICAL THINKING SKILLS INSTRUMENT AND DEEP LEARNING GRADUATES PROFILE IN SCIENCE LEARNING. *NUSRA: Jurnal Penelitian Dan Ilmu Pendidikan*, 7(4), 2330–2336. <https://doi.org/10.55681/nusra.v7i4.6653>



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

INTRODUCTION

The quality of measurement instruments is a fundamental prerequisite for valid and accountable educational research. Invalid and unreliable instruments will produce misleading data, making research conclusions unreliable (Fraenkel et al., 2019). In research and development (R&D), instrument validation is a critical step that must be undertaken before empirical data collection.

This research is part of a dissertation entitled “Development of a Video-Assisted PBEDL Model to Improve Independence, Collaboration, Critical and Creative Thinking Skills in Science Learning”. The PBEDL model integrates Project-Based Learning, Ethnoscience, and Deep Learning in a cohesive pedagogical framework. To measure the effectiveness of this model, an instrument is needed that is able to measure two main dependent variables: (1) critical thinking skills, and (2) the profile of deep learning graduates which includes independence, collaboration, and creativity.

Critical thinking skills are fundamental 21st-century skills, encompassing the ability to systematically analyze, evaluate, and synthesize information (Ennis, 1987; Facione, 2011). Meanwhile, the profile of deep learning graduates refers to the characteristics of students who are capable of in-depth, independent, collaborative, and creative learning, as mandated by the Ministry of Education, Culture, Research, and Technology (2025). These two aspects are measurable targets in this study.

Instrument validation plays a vital role in educational research. Without proper validation, the collected data cannot provide an accurate picture of the phenomenon being

studied. Therefore, this study specifically describes the procedures and results of content validity and reliability tests for both instruments, so that they can serve as references for other researchers developing similar instruments.

THEORITICAL REVIEW

Critical Thinking Skills and Ennis' Taxonomy

Ennis (1987) defines critical thinking as reflective and rational thinking focused on deciding what to believe or do. Ennis's taxonomy identifies five skill groups: (1) Elementary Clarification—focusing questions, analyzing arguments; (2) Basic Support—assessing source credibility, observing; (3) Inference—deduction, induction, value judgments; (4) Advanced Clarification—defining terms, identifying assumptions; (5) Strategies and Tactics—deciding on actions, interacting. These five indicators are operationalized in a contextual essay test instrument on the theme of Earth Structure and Geological Dynamics.

Deep Learning Graduate Profile

Deep Learning in the context of education (Fullan & Langworthy, 2014) refers to a learning process that encourages deep understanding through active engagement, meaningful connections, and the transfer of knowledge to new situations. The Ministry of Education, Culture, Research, and Technology (2025) identified the profile of relevant deep learning graduates as including: (a) Independence—self-awareness, self-regulation, and intrinsic motivation; (b) Collaboration—teamwork, active communication, and mutual respect; (c) Creativity—the ability to produce

original ideas, innovative work, and flexibility of thought.

Content Validity: CVR and CVI

The Content Validity Ratio (CVR) was developed by Lawshe (1975) with the formula: $CVR = (ne - N/2) / (N/2)$, where ne is the number of validators who rated the item as essential and N is the total number of validators. For $N = 5$ validators, the minimum acceptable CVR value at $p < 0.05$ is 0.99. The Content Validity Index (CVI) is the average CVR of all items; a CVI value ≥ 0.80 indicates good content validity (Polit & Beck, 2006). Wilson et al. (2012) suggested a CVI value ≥ 0.78 as the minimum acceptable limit.

Reliability: Cohen's Kappa and Cronbach's Alpha

Cohen's Kappa (κ) measures the level of agreement between two raters beyond random chance. Landis and Koch (1977) classified it as: $\kappa < 0.20$ (poor), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (good), 0.81–1.00 (excellent). Cronbach's Alpha (α) measures the internal consistency of the instrument; an α value ≥ 0.70 is acceptable, $\alpha \geq 0.80$ is excellent, $\alpha \geq 0.90$ is excellent (George & Mallery, 2003).

METHOD

This study uses a descriptive-quantitative approach within an R&D framework. The validation procedure follows Borich's (1994) model modified for the context of science education instruments.

Validators and Respondents

Validation involved 5 expert validators with the following criteria: active lecturers/researchers in the field of science education or educational evaluation, at least 5 years of experience, and at least a doctoral degree. The reliability test of the KBK

instrument was conducted by 2 raters (experienced senior science teachers). The graduate profile questionnaire was piloted on 30 eighth-grade junior high school students selected by purposive sampling from schools with similar characteristics to the main research subjects.

Validated Instruments

(a) KBK Instrument: Essay test 3 discourses $\times$ 5 sub-questions = 15 items, each measuring one indicator of Ennis KBK. Scoring rubric scale 0–4, maximum score 60. Validation sheet contains 9 assessment items scaled 1–5 covering aspects of format, language, suitability of indicators, and accuracy of material.

(b) Graduate Profile Questionnaire: 27 statements (18 positive, 9 negative), 4-point Likert scale (1 = STS to 4 = SS), covering three dimensions @ 9 items (3 subdimensions $\times$ 2 positive + 1 negative). The validation sheet uses V (4), CV (3), KV (2), TV (1) scales.

Data analysis

Content validity was analyzed using CVR and CVI. The reliability of the KBK instrument was analyzed using Cohen's Kappa, while the graduate profile questionnaire used Cronbach's Alpha with negative item recoding. The validity categories of CVI were: very valid (> 0.90), valid (0.79–0.90), quite valid (0.70–0.78), and less valid (< 0.70).

RESULTS AND DISCUSSION

1. Content Validity of Critical Thinking Skills Instrument

Table 1 presents a recapitulation of the validator's assessment scores for the 9 items of the KBK instrument validation sheet along with the CVR value of each item.

Table 1. Results of the Content Validity Assessment of the Critical Thinking Skills Instrument

Assessment Items	V1	V2	V3	V4	V5	Average	CVR	Note:
Instrument format according to provisions	5	5	4	5	5	4.80	1,000	Valid
Filling instructions are understandable	4	5	5	4	5	4.60	1,000	Valid
The language used is correct and standard	5	4	5	5	4	4.60	1,000	Valid
Easy to understand questionnaire	5	5	4	4	5	4.60	1,000	Valid
Writing accuracy is met	4	5	5	5	4	4.60	1,000	Valid
Statement according to Ennis KBK indicators	5	5	5	4	5	4.80	1,000	Valid
The number of questions per indicator is sufficient	4	5	5	5	4	4.60	1,000	Valid
The concept and terms are correct	5	5	4	5	5	4.80	1,000	Valid
Questions according to the level of junior high school students	5	4	5	4	5	4.60	1,000	Valid
CVI (mean CVR)							1,000	Very Valid

Based on Table 1, all validator assessment items obtained a CVR value of 1.000. This means that all validators agreed that all aspects of the instrument were essential. The overall CVI value reached 1.000, far exceeding the minimum limit of 0.99 for $N = 5$. The validators gave an average score of 4.60–4.80 on a scale of 1–5, indicating that the instrument was considered very good in terms of format, language, suitability of Ennis indicators, and appropriateness of the material for junior high school students.

Table 2. Content Validation Results of the Deep Learning Graduate Profile Instrument

No	Statement Items	V1	V2	V3	V4	V5	Average	CVR	Note:
	A. Independence (Items 1-9)								
1	Doing tasks without waiting for orders (+)	V	V	V	V	V	4.00	1,000	Valid
2	Dare to try to answer new questions (+)	V	V	V	V	V	4.00	1,000	Valid
3	Waiting for a friend to explain rather than understanding it yourself (-)	V	V	CV	V	V	3.80	0.600	Quite Valid
4	Make a regular science study schedule (+)	V	V	V	V	V	4.00	1,000	Valid
5	Checking work before submission (+)	V	V	V	V	V	4.00	1,000	Valid
6	Postpone tasks until near the deadline (-)	CV	V	V	V	V	3.80	0.600	Quite Valid
7	Studying science because I want to understand the material (+)	V	CV	V	V	V	3.80	0.600	Quite Valid
8	Searching for additional learning resources independently (+)	V	V	V	V	V	4.00	1,000	Valid
9	Lazy to study without supervision or reminders (-)	V	V	V	CV	V	3.80	0.600	Quite Valid

	B. Collaboration (Items 10-18)								
10	Actively contribute to group activities (+)	V	V	V	V	V	4.00	1,000	Valid
11	Willing to share tasks fairly (+)	V	V	V	V	V	4.00	1,000	Valid
12	Let friends do all the group work (-)	V	CV	V	V	V	3.80	0.600	Quite Valid
13	Listening to friends' opinions during discussions (+)	V	V	V	V	V	4.00	1,000	Valid
14	Convey ideas clearly during discussions (+)	V	V	V	V	V	4.00	1,000	Valid
15	Ignore different opinions of friends (-)	V	V	V	V	CV	3.80	0.600	Quite Valid
16	Encourage friends who are having difficulties (+)	V	V	V	V	V	4.00	1,000	Valid
17	Respect differences of opinion between members (+)	V	V	V	V	V	4.00	1,000	Valid
18	Feeling that your own opinion is always the most correct (-)	V	V	CV	V	V	3.80	0.600	Quite Valid

	C. Creativity (Items 19-27)								
19	Dare to propose new ideas during discussions (+)	V	V	V	V	V	4.00	1,000	Valid
20	Trying different ways of solving problems (+)	V	V	V	V	V	4.00	1,000	Valid
21	Prefer to follow existing methods (-)	CV	V	V	V	V	3.80	0.600	Quite Valid
22	Create reports with a unique appearance (+)	V	V	V	V	V	4.00	1,000	Valid
23	Utilizing surrounding materials for innovative work (+)	V	V	V	V	V	4.00	1,000	Valid
24	Prefer to copy friends' work (-)	V	V	V	CV	V	3.80	0.600	Quite Valid
25	Looking for another way when the experiment fails (+)	V	V	V	V	V	4.00	1,000	Valid
26	Thinking of more than one solution (+)	V	V	V	V	V	4.00	1,000	Valid
27	Give up if the first method doesn't work (-)	V	CV	V	V	V	3.80	0.600	Quite Valid
	CVI (average CVR 27 items)							0.852	Valid

Of the 27 graduate profile instrument items, 17 items obtained a CVR of 1.000 (Valid) and 10 items obtained a CVR of 0.600 (Quite Valid). The overall CVI value of 0.852 is in the valid category ($CVI > 0.79$). Items with a CVR value of 0.600 are generally negative items (unfavorable statements). Based on the validator's notes, these items are still declared usable with minor revisions to the clarity of the wording

3. Reliability Test of Critical Thinking Skills Instrument (Inter-Rater Reliability)

The reliability test of the KBK instrument was conducted by calculating the coefficient of agreement (Cohen's Kappa) between two raters who assessed 30 student answer sheets on 15 sub-questions. Table 3 presents the results per Critical Thinking Skills indicator

Table 3. Results of the Inter-Rater Reliability Test of the critical thinking skills Instrument (Cohen's Kappa)

No	critical thinking skills Indicator (Ennis)	Rater 1	Rater 2	κ (Kappa)	Category
1	Elementary Clarification	2.73	2.80	0.821	Very good
2	Basic Support	2.67	2.73	0.813	Very good
3	Inference (Concluding)	2.80	2.87	0.847	Very good
4	Advanced Clarification	2.63	2.70	0.808	Very good
5	Strategies and Tactics	2.77	2.83	0.839	Very good
	Overall Average	2.72	2.79	0.826	Very Reliable

The Kappa coefficient for each indicator of Critical Thinking Skills ranges from 0.808 to 0.847, with an average κ of 0.826. All κ values fall into the 'very good' category (almost perfect agreement). This means that both raters have a very high agreement in using the scoring rubric of the CBC instrument, so the instrument can be declared reliable for inter-rater assessment.

4. Reliability Test of Graduate Profile Questionnaire (Cronbach's Alpha)

Table 4 presents the results of the Cronbach's Alpha analysis of the graduate profile questionnaire trial data from 30 respondents, after recoding negative items.

Table 4. Cronbach's Alpha Reliability Test Results for the Deep Learning Graduate Profile Questionnaire

No	Dimensions	Σ Item	Σ Var. Item	Alpha (α)	Category
1	Independence	9	0.324	0.828	Very good
2	Collaboration	9	0.317	0.841	Very good
3	Creativity	9	0.331	0.818	Very good
	Total Instruments (27 items)	27	0.972	0.946	Very good

Alpha Coefficient The Cronbach's α per dimension ranged from 0.818 to 0.841, and the total alpha reached 0.946. All these values fall within the 'very good' category ($\alpha \geq 0.80$). High internal consistency indicates that the items within each dimension adequately measure the same construct.

The high total alpha also indicates that the three dimensions correlate with each other in holistically measuring the profile of deep learning graduates.

5. Recapitulation and Discussion

Table 5. Summary of Validity and Reliability Results of the Two Instruments

Instrument	Number of Items	CVI	Reliability	Decision
KBK (Essay Test) Instrument	15 sub-questions	1.000 (Very Valid)	$\kappa = 0.826$ (Very Reliable)	Worth Using
DL Graduate Profile Questionnaire	27 statements	0.852 (Valid)	$\alpha = 0.946$ (Very Reliable)	Worth Using*

The results of this study indicate that both instruments developed have met the required psychometric standards. The KBK instrument, with a CVI of 1.000, exceeds the minimum CVI recommended by Wilson et al. (2012), which is 0.78, and even approaches a perfect value. This reflects the quality of the instrument development, which is comprehensively based on Ennis's (1987) theory and validated by relevant experts.

The graduate profile instrument, with a CVI of 0.852, is considered valid. This is understandable, considering that the constructs of "independence," "collaboration," and "creativity" are multidimensional and more abstract than the KBK construct, which has an established taxonomy. Negative items rated "CV" by several validators reflect the general challenges in developing unfavorable items in affective instruments.

Excellent reliability for both instruments ($\kappa = 0.826$ and $\alpha = 0.946$) confirms that the instruments produce consistent and reliable data. A high Kappa coefficient indicates that the scoring rubric

has been formulated explicitly enough to reduce subjectivity in assessment. A high Cronbach's alpha indicates good internal homogeneity, indicating that the items measure the same construct coherently.

CONCLUSION

Based on the analysis results, it can be concluded that: (1) The Critical Thinking Skills Instrument based on Ennis' taxonomy is declared very valid (CVI = 1.000) and very reliable ($\kappa = 0.826$), so that all 9 validation items and 15 sub-

questions are suitable for use without revision. (2) The Deep Learning Graduate Profile Questionnaire (27 items) is declared valid (CVI = 0.852) and very reliable ($\alpha = 0.946$). A total of 17 items are declared valid and 10 items are declared quite valid with suggestions for minor revisions to the clarity of the wording. (3) Both instruments are suitable for use in research on the development of a video-assisted PBEDL model to measure critical thinking skills and profiles of deep learning graduates of junior high school students in science learning. The researcher suggests: (a) conducting a trial with a larger sample ($n \geq 100$) for exploratory factor analysis; (b) revising the wording of the 10 items that received a CVR value of 0.600 before being used in the main research; (c) consideration of adding construct validity tests using CFA (Confirmatory Factor Analysis) to verify the three-factor structure of the graduate profile instrument.

REFERENCE

- Borich, G. D. (1994). *Observation skills for effective teaching*. Macmillan.
- Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In JB Baron & RJ Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9–26). WH Freeman.
- Facione, P. A. (2011). *Critical thinking: What it is and why it counts*. Insight Assessment.
- Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2019). *How to design and evaluate research in education* (10th ed.). McGraw-Hill Education.
- Fullan, M., & Langworthy, M. (2014). *A rich seam: How new pedagogies discover deep learning*. Pearson.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.
- Ministry of Education, Culture, Research, and Technology. (2025). *Elementary and Secondary Education Graduate Profile Framework*. Ministry of Education, Culture, Research, and Technology.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497. <https://doi.org/10.1002/nur.20147>
- Wilson, F.R., Pan, W., & Schumsky, D.A. (2012). *Recalculation of the critical*

- values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45(3), 197–210. <https://doi.org/10.1177/0748175612440286>
- Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, 4(2), 165–178. <https://doi.org/10.15171/jcs.2015.017>
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
- Azwar, S. (2012). *Reliability and validity* (4th ed.). Student Library.
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson Education.
- Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills: Methods and approaches*. Springer. <https://doi.org/10.1007/978-94-017-9395-7>
- Hair, JF, Black, WC, Babin, BJ, & Anderson, RE (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.
- Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, 34(3), 177–189.
- 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>
- Jonassen, D.H. (2011). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge.
- Ministry of Education, Culture, Research, and Technology. (2022). Decree of the Minister of Education, Culture, Research, and Technology No. 56/M/2022 concerning Guidelines for Curriculum Implementation in the Context of Learning Recovery. Ministry of Education, Culture, Research, and Technology.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Sugiyono. (2019). *Quantitative, qualitative, and R&D research methods* (2nd ed.). Alfabeta.