Chemined 15 (1) (2026)

Chemistry in Education

https://journal.unnes.ac.id/journals/chemined

Analysis of Students' Critical Thinking Skills in Electrochemistry Through Artificial Intelligence (AI)-Assisted Video Storytelling

Nor Unsa Akmal[1*)] and Harjono[1]

[1]Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

ARTICLE INFO

Article history:
Received: February 2026
Approved: July 2026
Published: July 2026

Kata Kunci:
Kemampuan Berpikir Kritis, Elektrokimia, Video Storytelling, Kecerdasan Buatan.

Keywords:
Critical Thinking Skills, Electrochemistry, Video Storytelling, Artificial Intelligence

DOI:
https://doi.org/10.15294/chemined.v15i1.43619

ABSTRAK

Tuntutan pembelajaran abad ke-21 yang menekankan keterampilan berpikir kritis sebagai bagian dari 4C, sehingga diperlukan instrumen penilaian yang mampu mengungkap proses berpikir siswa secara lebih autentik pada materi elektrokimia. Berdasarkan penelitian yang dilaksanakan di SMAN 3 Semarang pada November 2025 dengan melibatkan 86 siswa kelas XII menunjukkan bahwa penilain masih bersifat tradisional menggunakan tes tulis berbasis kertas dan penilaian masih bersifat subjektif. Metode kuantitatif dan desain korelasional yang digunakan pada penelitian ini berfokus pada pengujian tes storytelling berbantuan dengan cara divideo kemudian transkrip dilakukan dengan bantuan kecerdasan buatan (AI) untuk mengukur kemampuan berpikir kritis berdasarkan indikator Facione (interpretasi, analisis, inferensi, evaluasi, dan eksplanasi). Kelayakan instrumen ditinjau melalui validitas isi oleh dua validator dan dianalisis menggunakan rumus Gregory, serta reliabilitas diuji menggunakan Cronbach's Alpha. Hasil menunjukkan koefisien validitas isi sebesar 0,80 (kategori sangat tinggi) sehingga instrumen dinilai valid secara isi,. Reliabilitas instrumen memperoleh α = 0,623 dikatakan cukup reliabel. Profil kemampuan berpikir kritis siswa cenderung tinggi pada Interpretasi–Analisis–Inferensi, sementara Evaluasi dan Eksplanasi relatif lebih rendah. Secara kategorikal berdasarkan skor total, sebesar 88,37% siswa berada pada kategori sangat kritis, 10,47% kategori kritis, dan 1,16% kategori cukup kritis. Meskipun hasil menunjukkan bahwa hubungan antar indikator KBK cenderung lemah.

ABSTRACT

The demands of 21st-century learning emphasize critical thinking skills as part of the 4Cs, requiring assessment instruments capable of revealing students' thinking processes more authentically in electrochemistry material. Based on research conducted at SMAN 3 Semarang in November 2025 involving 86 grade XII students, it was found that assessment was still traditional, using paper-based written tests, and was still subjective. The quantitative method and correlational design used in this study focused on testing assisted storytelling tests by means of video recording and transcription with the help of artificial intelligence (AI) to measure critical thinking skills based on Facione's indicators (interpretation, analysis, inference, evaluation, and explanation). The instrument's feasibility was reviewed through content validity by two validators and analyzed using Gregory's formula, while reliability was tested using Cronbach's Alpha. The results showed a content validity coefficient of 0.80 (very high category), indicating that the instrument was valid in terms of content. The instrument's reliability obtained an α = 0.623, which is considered sufficiently reliable. Students' critical thinking skills tend to be high in Interpretation–Analysis–Inference, while Evaluation and Explanation are relatively lower. Categorically, based on total scores, 88.37% of students are in the very critical category, 10.47% in the critical category, and 1.16% in the fairly critical category. Although the results show that the relationship between Critical Thinking indicators tends to be weak.

*) **Correspondence address:**
Jl. Durian Sekaran, Gg. Rambutan, Kec. Gunung Pati, Kota Semarang
E-mail: norunsaakmal98@gmail.com

p-ISSN 2252-6609
e-ISSN 2502-6852

INTRODUCTION

Twenty-first-century learning foregrounds the 4C competencies critical thinking, creativity, communication, and collaboration as essential capacities for students to navigate global challenges and interdisciplinary demands (Nurhayati et al., 2024). In Indonesia, these priorities are reflected in Curriculum 2013 and Kurikulum Merdeka, which emphasise character formation and competence development, particularly critical and creative thinking (Imanda et al., 2024). Kurikulum Merdeka also advances deep learning that balances academic achievement with character, creativity, and empathy (Puspa et al., 2023), yet its effectiveness depends on teachers' readiness and sustained competence-based practice (Suherman et al., 2022). Consequently, strengthening critical thinking is both an educational mandate and a strategic need for preparing learners for Society 5.0.

Critical thinking in chemsitry is indispensable because students must interpret evidence, analyse mechanisms, make justified inferences, and explain relationships across representations rather than merely recall facts (Hanania et al., 2022). These demands are particularly pronounced in electrochemistry, where abstract concepts such as electron transfer and cell potential require systematic higher-order reasoning and conceptual integration (Szozda et al., 2022). This study adopts Facione's critical thinking framework operationalised here through interpretation, analysis, inference, evaluation, and explanation (Facione, 2020) as a robust basis for examining how students construct and justify understanding in electrochemistry, consistent with evidence that critical thinking supports chemistry learning outcomes (Wahyudiati, 2022).

Assessing critical thinking remains challenging when conventional formats (e.g., multiple-choice items or short, constrained essays) predominantly capture recall and limit visibility of reasoning processes (Radiansyah et al., 2024). More open and contextual assessment approaches are argued to better elicit higher-order thinking because they require students to explore, justify, and solve problems in authentic situations (Dewi & Pramartha, 2021). Video storytelling is increasingly recognised as a pedagogical and assessment-relevant modality that can prompt students to organise ideas coherently, communicate arguments, and connect concepts to real-world contexts (e.g., corrosion and batteries), thereby eliciting richer evidence of reasoning than traditional tests (Huang et al., 2020).

The analytic yet workload of evaluating large volumes of verbalised student reasoning from videos can be prohibitive without technological support. Advances in AI and NLP enable scalable processing of educational data (e.g., transcripts) and offer the potential for more efficient, evidence-based, and less subjective assessment (Suharyo et al., 2024), supported by cloud-based workflows (Aditya, 2021). AI can also detect patterns and relationships within students' reasoning (e.g., via clustering) and can be evaluated using approaches such as Pearson correlation and confusion matrices when validating classification or scoring accuracy (Zen et al., 2023). Despite this promise, explicit integration of video

storytelling, AI-assisted transcription/scoring, and Facione-indicator analysis in electrochemistry especially in Indonesian senior secondary contexts remains limited, creating a clear research gap.

Therefore, this study proposes an innovative assessment approach by using students' electrochemistry video storytelling responses and employing AI tools (chem.ai for transcription and Gemini AI Flash 3.0 for analysis/scoring) to examine critical thinking across Facione-aligned indicators (Facione, 2020). Conducted with Year 12 students at SMA Negeri 3 Semarang in a non-interventional design, the study addresses three research questions: (1) What are the validity and reliability characteristics of a storytelling-based test instrument for measuring Facione-indicated critical thinking in electrochemistry? (2) What is the profile of students' critical thinking across interpretation, analysis, inference, evaluation, and explanation? and (3) What relationships exist among these critical thinking indicators based on students' video storytelling data? Accordingly, the study aims to establish the instrument's psychometric quality, describe students' critical thinking profiles in electrochemistry, and analyse inter-indicator relationships to inform more contextual, technology-enhanced assessment practices aligned with twenty-first-century learning.

METHODS

Reasearch design and procedure

This study employed a quantitative approach to measure students' critical thinking skills and to evaluate the accuracy of AI-assisted scoring. A quantitative design was selected because it enables numerical representation of students' performance, statistical testing of relationships among constructs, and measurable validation of AI-generated assessment outcomes in a way that supports broader inference (Creswell, J. W., & Plano Clark, 2023). In line with this purpose, a correlational design was used to examine the relationships among critical thinking dimensions operationalised from Facione's indicators (interpretation, analysis, inference, evaluation, and explanation).

The research workflow implemented in this study is illustrated in **Figure 1**. The research procedure began with test instrument development. An essay-based test was designed to elicit evidence of critical thinking in the electrochemistry topic area, with each prompt constructed to align explicitly with Facione's indicators (interpretation, analysis, inference, evaluation, and explanation). The draft instrument then underwent expert validation to establish content validity, focusing on the alignment between items and research objectives, clarity and readability for Grade 12 students, and consistency with the applicable curriculum. Feedback from the experts was used to refine the items to ensure the instrument was fit for data collection.

Following validation data were collected from students who had completed learning on electrochemistry. Students responded to the essay prompts by recording their answers as video storytelling presentations (maximum duration 10 minutes) using their own cameras. To reduce potential

bias arising from background noise and voice overlap, recordings were organised in shifts within each class. Students uploaded their videos to YouTube, activated subtitles, and submitted the video links via a Google Form. The collected responses were then transcribed for analysis. In the scoring stage, students' transcribed answers were evaluated using a Facione-indicator rubric, with scoring supported by the Gemini API Flash 3.0 to generate systematic quantitative scores for each indicator. These scores constituted the primary dataset for subsequent statistical analysis, including examination of inter-indicator relationships and evaluation of AI-assisted scoring performance.

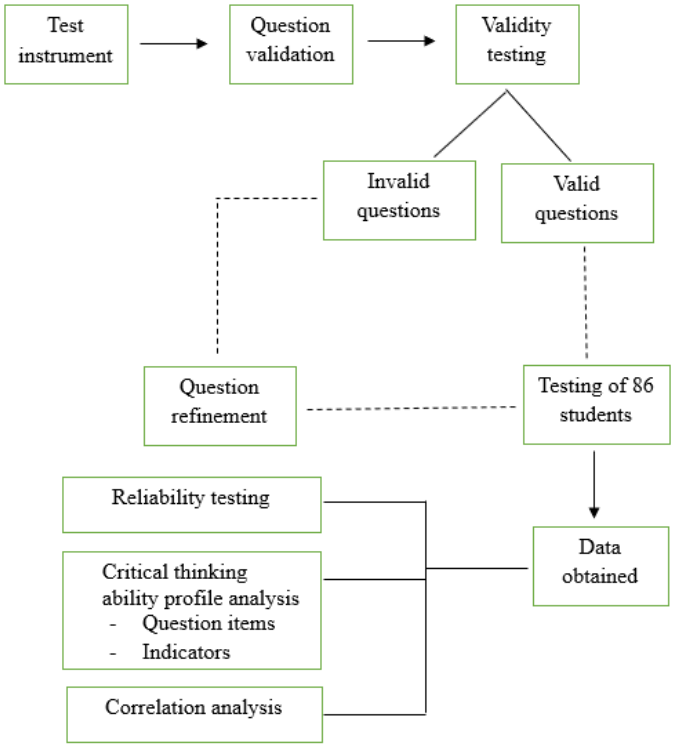

Figure 1. Reasearch Process Flow

Data and instrumentation

Data were obtained from multiple sources aligned with each research activity: (i) a preliminary study to map the context of Grade 12 electrochemistry learning (books, journal articles, reputable online resources, school stakeholders/administration, and students), (ii) expert validation data to establish the content validity of the critical thinking instrument (expert validators), and (iii) student performance data from the storytelling-based test to capture critical thinking attainment (students). In line with the principle that data collection is central to achieving research objectives, the study prioritised procedures that generated analysable and accountable evidence (Sugiyono, 2019).

The primary instrument was a storytelling-based critical thinking test administered as an oral video response rather than a written paper-and-pencil format. The test comprised two essay items, each containing five sub-questions designed to elicit students' reasoning in electrochemistry across Facione-aligned indicators (interpretation, analysis, inference, evaluation, and explanation). The test was

administered in a controlled classroom setting and divided into three shifts to minimise noise and recording bias. Students worked under regular-classroom conditions with a total allocation of 30 minutes per shift (including reading, analysing, and responding). Each essay item was recorded as a separate video (thus, two videos per student). Students recorded responses in front of their cameras, uploaded the videos to YouTube, enabled subtitles, and submitted the links via a Google Form. To standardise conditions and reduce external interference, the process was supervised by the subject teacher and the research team, and students were not permitted to access textbooks or digital sources during the assessment.

To ensure data trustworthiness, the instrument was evaluated through validity and reliability testing. Content validity was established through expert judgement by a university lecturer and a chemistry teacher, who reviewed each item using three criteria: alignment with the intended critical thinking indicator, clarity of wording, and appropriateness for senior secondary cognitive level. Item ratings followed a 4-point scale (1–2 = not relevant; 3–4 = relevant), and inter-expert agreement was quantified using the Gregory content validity index, providing a numerical basis for interpreting content validity strength (Sugiarta et al., 2024). Reliability was assessed via internal consistency using Cronbach's alpha, which is widely used for educational instruments to evaluate consistency across related items (Beyera et al., 2020). In this study, scoring was supported by Gemini AI (Gemini API Flash 3.0) using a predefined rubric; the AI was required to provide a justification for each score to enhance transparency. Where the AI's rationale did not align with the rubric criteria or the targeted indicator, the transcript was re-evaluated and rescored to maintain consistency with the intended measurement construct.

Descriptive analysis

Descriptive statistical analysis in this study was used to summarise and present students' critical thinking data as it is, without hypothesis testing, which is common in educational research when the aim is to portray learners' achievement patterns (Susilawati et al., 2020). The analysis reports measures of central tendency (mean) and dispersion (SD) so that readers can understand not only the average level of performance but also the variability and distribution of students' abilities (Susilawati et al., 2020). Because critical thinking was operationalised through predetermined indicators (referring to Facione), descriptive profiling by indicator is essential to reveal which dimensions are relatively stronger and which require further reinforcement, consistent with prior profiling studies in different learning contexts (Rosidah et al., 2021). In addition, presenting scores as percentages of the maximum possible score allows clearer interpretation of how close students' achievement is to an ideal benchmark for each indicator (Erwanto, 2020). This indicator-level profile is particularly relevant in digital-media-based learning contexts such as video learning or digital storytelling, where multimedia formats may stimulate higher-order thinking differently across dimensions; thus, descriptive results can highlight which indicators are most supported by the medium and which remain challenging for students (Zuhannisa' et al., 2023).

Correlation analysis

Inter-indicator correlation analysis in this study employed Spearman's rank correlation to examine the direction and strength of relationships among the five Facione-aligned critical thinking indicators—interpretation, analysis, inference, evaluation, and explanation—which are widely adopted as core components for developing assessment indicators and rubrics (Astuti et al., 2022). Spearman correlation was selected because critical thinking in educational research is commonly operationalised through essay tasks scored with rubrics using limited rating scales (e.g., 1–5), producing ordinal (ranked) data; thus, a non-parametric, rank-based approach is more appropriate than parametric correlations as it does not require normally distributed data and fits ordinal measurement characteristics (Kamalia & Wasis, 2021). Spearman analysis is also frequently used in education to interpret linkages among critical thinking components and related competencies, strengthening integrated interpretations of students' profiles (Kristianawati et al., 2022). In this study, the correlation results indicate whether achievement in one indicator tends to co-vary with others—suggesting a coherent critical thinking profile—or whether weaker relationships signal uneven development across indicators, thereby informing targeted instructional support. Additionally, because some sub-items were designed to measure the same indicator across different prompts, correlation analysis provides supplementary evidence of alignment across similar items, supporting the claim that rubric-based scores represent the same underlying construct conceptually (Kamalia & Wasis, 2021).

RESULT AND DISCUSSION

Result

Validity and Reliability Test Results for Test Instruments

Test Instrument Validity Test Results

Content validity testing was conducted to ensure that each storytelling test item was relevant to the intended critical thinking indicators and appropriate for the electrochemistry context.

Table 1. Result of Gregory's Content Validity Test

Question Number	Validator 1	Validator 2	Tabulation
1a	3	4	D
1b	2	4	C
1c	2	3	C
1d	4	3	D
1e	4	3	D
2a	3	4	D
2b	4	4	D
2c	4	3	D
2d	4	3	D
2e	4	3	D

The review involved two validators—Validator 1 (a lecturer with expertise in instrument development) and Validator 2 (a senior secondary chemistry teacher)—and the results were analysed using the Gregory formula, as summarised in Table 1. The Gregory coefficient obtained was 0.80, indicating high to very high content validity according to the validity categories. Although Validator 1 noted that two sub-items (1b and 1c) were less relevant, Validator 2 rated most items as relevant; overall, the calculated Gregory index confirms that the instrument is highly valid and suitable for measuring students' critical thinking in electrochemistry.

Reliability Test Results

Instrument reliability was assessed to examine internal consistency across items, that is, the extent to which the questions function coherently in measuring the same construct. Using SPSS on data from 86 students, Cronbach's Alpha yielded α=0.623, as presented in Table 2.

Table 2. Cronbach's Alpha Reliability Test Results

Cronbach's Alpha	N	Conclusion
0,623	10	$0{,}6 < \alpha < 0{,}7$ (quite reliable)

Indicating quite reliable/acceptable level of reliability according to the applied criteria. Conceptually, Cronbach's Alpha reflects the internal consistency of an instrument, where higher values indicate more dependable measurement (Rosli et al., 2021) therefore, α = 0.623 suggests the test is sufficiently reliable for use in this study, while also indicating that further refinement could improve reliability in future applications.

Critical Thinking Ability Profile Analysis Test Results

Descriptive Statistical Analysis Results per Item

Item-level descriptive statistics (Items 1a–2e) were generated using SPSS to summarise students' score characteristics for each item in a concise and systematic manner; the use of SPSS for presenting measures of central tendency and dispersion at the item level is common in educational instrument analysis because it helps describe both attainment and response variability across items (Mutakin, 2023).

As shown in Table 3, complete data were obtained from all 86 students, indicating that the descriptive results represent the full sample. The mean scores across the ten sub-items were high (4.30–4.97), with Item 1a showing the highest mean (4.97), suggesting that most students demonstrated very strong performance on interpretation, whereas the lowest mean occurred on Item 1d, indicating comparatively greater difficulty and variation in evaluation performance. Standard deviations in Table 3 indicate the extent of score variability, where smaller SD values reflect more homogeneous responses and larger SD values reflect greater diversity in students' answers (Wahyuni et al., 2022); the largest SD values were observed for Items 1e and 1d (greater variation in explanation and evaluation), while the smallest SD values appeared for Items 2c and 1a (more consistent responses for inference and

interpretation). All items had a maximum score of 5, showing that at least some students achieved the highest possible performance on each item (Susilaningsih & Mudjijanti, 2024), whereas minimum scores varied across items, with several items (1b, 1d, 1e, 2d, and 2e) reaching a minimum of 1, implying wide performance gaps and heterogeneous understanding for those indicators; other minima were 2 (Item 1c), 3 (e.g., Item 1a), and 4 (Items 2b and 2c). To classify overall critical thinking level, student scores were converted into percentages of the maximum score, and the distribution of categories is reported in Table 4.

Table 3. Results of Descriptive Statistical Analysis of Individual Questions

Question Number	N	Mean	SD	Minimum	Maximum
No 1a	86	4,97	0,24	3	5
No 1b	86	4,86	0,58	1	5
No 1c	86	4,78	0,82	2	5
No 1d	86	4,30	1,03	1	5
No 1e	86	4,31	1,15	1	5
No 2a	86	4,74	0,49	3	5
No 2b	86	4,76	0,43	4	5
No 2c	86	4,95	0,21	4	5
No 2d	86	4,52	0,76	1	5
No 2e	86	4,79	0,59	1	5

Table 4. Percentage of Students' Critical Thinking Test Results

Category	Interval (%)	Total score range (max 50)	Number of students	Percentage
Very Critical	86–100	43–50	**76**	**88,37%**
Critical	76–85	38–42	9	10,47%
Sufficient	60–75	30–37	1	1,16%
Insufficient	<60	0–29	0	0%

Descriptive Statistical Analysis Results per Critical Thinking Ability Indicator

Descriptive statistics by critical thinking indicator were used to provide a clearer, more comparable overview of students' achievement across Facione's dimensions (interpretation, analysis, inference, evaluation, and explanation) than item-level scores alone. Mean scores indicate students' average mastery for each indicator, while standard deviations (SD) reflect the extent of variation among students (Chamberlain et al., 2021); presenting results as a percentage of the maximum score further clarifies how close performance is to the ideal benchmark (Septiawati & Eftanastarini, 2020).

Table 5. Results of Descriptive Statistical Analysis of CTA Indicators

Indicator	N	Mean	SD	% from Max
Interpretation	86	4.68	0.72	93.6%
Analysis	86	4.61	0.81	92.2%
Inference	86	4.63	0.73	92.6%
Evaluation	86	4.12	1.00	82.4%
Explanation	86	4.30	0.96	86.0%

As shown in Table 5, all indicators achieved high mean scores approaching the maximum of 5, although score dispersion differed across indicators. Interpretation recorded the highest mean, whereas evaluation was the lowest and showed the greatest variability, followed by explanation; the most consistent performance (lowest variability) occurred in interpretation and inference. Overall, the percentage achievements were highest for interpretation (93.6%), followed by inference (92.6%), analysis (92.2%), explanation (86.0%), and evaluation (82.4%).

Correlation Analysis Results between Critical Thinking Indicators

Spearman correlation across the ten sub-items (Scores 1a–1e and 2a–2e) was used to examine the strength and direction of inter-item relationships under ordinal scoring and/or non-parametric conditions (Casinillo, 2022). In the heatmap (Figure 2), each cell represents Spearman's rho (−1 to +1), where positive values indicate scores tend to increase together and negative values indicate an inverse tendency (Casinillo, 2022); such inter-item correlation analysis is commonly reported in psychometric evaluation to reflect item relatedness and construct cohesion alongside reliability and construct validity evidence (Dolansky et al., 2020).

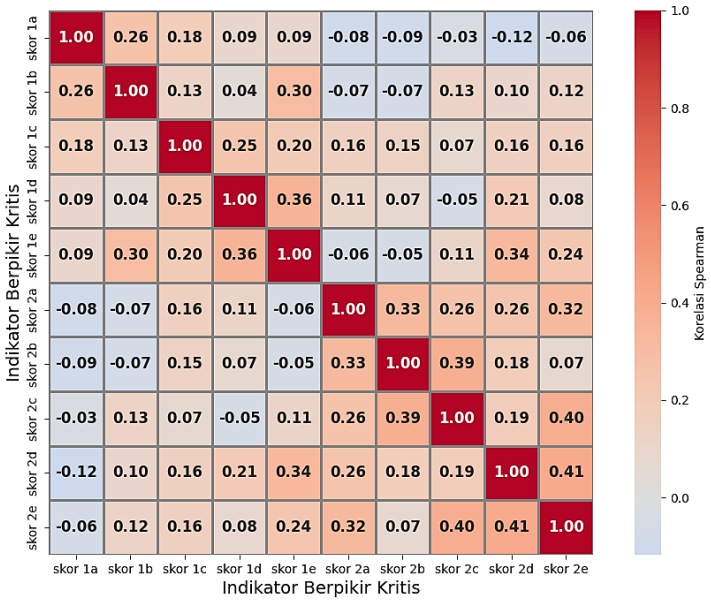

Figure 2. Heatmap Graph of Spearman Correlation Analysis Results

Figure 2 shows mostly weak correlations (approximately −0.12 to 0.30), suggesting that performance on one sub-item does not strongly track performance on others; the few negative correlations were trivial (e.g., 1a–2d, ρ = −0.12; 1a–2b, ρ = −0.09). Nevertheless, several pairs were more notable, including 2d–2e (ρ = 0.41), 2c–2e (ρ = 0.40), 2b–2c (ρ = 0.39), 1d–1e (ρ = 0.36), and 1e–2d (ρ = 0.34), indicating greater overlap in skill demands for these items. By contrast, "parallel" pairs intended to tap similar indicators across contexts (e.g., 1a–2a, 1b–2b, 1c–2c) showed low correlations (ρ = −0.08; −0.07; 0.07, respectively), implying that changes in item context may substantially shift students' response patterns even when the targeted indicator is nominally the same.

Discussion

The instrument analysis was conducted to address the study aims: establishing the validity and reliability of the electrochemistry storytelling test, profiling Grade 12 students' critical thinking (Facione indicators), and examining relationships among indicators using students' video storytelling data. Content validity was examined through expert judgement involving two validators (Validator 1: instrument expert lecturer; Validator 2: senior secondary chemistry teacher) and quantified using the Gregory coefficient. In the Gregory framework, category D reflects strong agreement that an item is relevant to the intended construct, whereas category C occurs when Validator 1 assigns 1–2 and Validator 2 assigns 3–4, indicating disagreement in perceived relevance for particular items (Nuryuliana & Prodjosantoso, 2021). The resulting Gregory index was 0.80, which indicates high to very high content validity and supports the instrument's suitability for measuring Facione-aligned critical thinking in electrochemistry (Suyasa & Divayana, 2022). Although disagreement occurred on a small number of sub-items (e.g., items 1b and 1c), such differences are commonly attributed to distinct professional lenses—instrument specialists may prioritise construct purity and language precision, whereas teachers may emphasise classroom feasibility and curriculum familiarity—suggesting that minor revisions (e.g., simplifying wording without altering the indicator demand and ensuring electrochemistry terms are unambiguous) could further strengthen item clarity while retaining construct alignment (Suyasa & Divayana, 2022). This magnitude of content validity (≈0.8) is also consistent with prior evidence that a coefficient around 0.8 reflects strong content foundations for measuring critical thinking (Eliza et al., 2022).

Internal consistency reliability was then assessed for the 10 sub-items (two question × five sub-questions) using Cronbach's alpha in SPSS with N = 86, yielding α = 0.623 (Table 2). Conceptually, Cronbach's alpha reflects internal consistency—higher values indicate more dependable measurement across items intended to assess the same general construct (Rosli et al., 2021)—and α values above 0.60 are often considered acceptable, especially in early-stage educational instrument development or relatively short instruments (Azhar et al., 2018). The "quite reliable" (rather than strong) alpha suggests

room for refinement. One plausible explanation is a ceiling effect: when many students obtain near-maximum scores, score variance shrinks, which can reduce inter-item covariance and depress alpha even if the construct is coherent (Rivas et al., 2023). This interpretation is consistent with the observed descriptive pattern of very high means across multiple sub-items. In addition, because scoring was supported by Gemini AI, overly generous scoring tendencies (e.g., rewarding fluency/structure more than scientific justification) could further concentrate scores at the top end; prior studies similarly caution that AI scoring may inflate scores based on linguistic smoothness rather than depth of reasoning (Tleshova et al., 2025). To mitigate this, this study required AI-generated scoring rationales and conducted human verification when rationales did not match rubric criteria; however, because formal inter-rater reliability between AI and human raters was not computed, future work should consider explicit rater-agreement analyses and stronger human moderation to reduce potential AI bias (Gencer & Doğan, 2020).

Item-level descriptive results (Table 3) further clarify why reliability and correlations may appear modest. Across all 10 sub-items, the means were high (4.30–4.97 out of 5), indicating that most students performed at a very high level on the storytelling tasks, with the highest mean on Item 1a (M = 4.97) and the lowest mean on Item 1d (M = 4.30). Standard deviations varied meaningfully across items: items such as 1a and 2c showed very small SDs (more homogeneous, high-end performance), whereas items such as 1d and 1e showed the largest SDs (greater spread), suggesting that evaluation and explanation demands differentiated students more strongly. Minimum scores also varied (some items reaching 1), showing that a small subset of students struggled substantially on particular sub-tasks even when overall performance was high. This pattern—high overall attainment with uneven difficulty across critical thinking components—aligns with findings that Facione-based profiles often show non-uniform performance, particularly for indicators requiring stronger justification (Gladly et al., 2024). When total scores were converted to percentage categories (Table 4), 88.37% of students were classified as very critical, 10.47% as critical, and 1.16% as sufficient, reinforcing the interpretation of generally high performance but also raising the possibility that some items were insufficiently discriminating at the top end (ceiling effect), which can limit the instrument's ability to separate high-performing students (Rivas et al., 2023).

Indicator-level descriptive analysis (Table 5) provides a clearer profile of students' critical thinking strengths and weaknesses in electrochemistry. Achievement was highest on Interpretation (93.6%), followed closely by Inference (92.6%) and Analysis (92.2%), while Evaluation (82.4%) was lowest and showed the greatest variability (highest SD), with Explanation (86.0%) also comparatively lower and variable. This hierarchy is consistent with prior studies reporting interpretation as a dominant strength and evaluation as a common difficulty, because evaluation requires explicit criteria, evidence-based judgement, and deeper scientific justification rather than recognition or summarisation (Sunarti et

al., 2023) Similarly, interpretation–analysis–inference often appear stronger because they are closer to comprehension and basic reasoning steps, whereas evaluation–explanation demand more explicit argumentation and elaboration (Ulfa & Permatasari, 2023). Evidence from transcripts also supported this pattern: high-scoring evaluation/explanation responses tended to include explicit electrochemistry principles and causal reasoning, whereas lower responses were more generic and lacked criteria or scientific warrants—consistent with the view that argumentation-focused tasks and prompts are necessary to elicit strong evaluative and explanatory reasoning (Rahmawati & Hidayati, 2023). This suggests that future revisions should prioritise improving the discriminative power and clarity of items targeting evaluation and explanation (e.g., items 1d, 2d, 2e).

The Spearman correlation heatmap (Figure 2) showed predominantly weak inter-item correlations (approximately −0.12 to 0.30), with only a few pairs reaching weak–moderate levels (e.g., 2d–2e = 0.41; 2c–2e = 0.40; 2b–2c = 0.39; 1d–1e = 0.36). Such a pattern is defensible for a multidimensional construct like critical thinking, where sub-items intentionally target different Facione indicators; hence, low correlations do not necessarily indicate poor items but may reflect limited overlap by design (Semegni et al., 2021). The relatively stronger correlations clustered among cognitively adjacent demands (notably evaluation–explanation pairs), which is theoretically plausible because stronger evaluative judgement often underpins coherent explanation (Salter et al., 2020). However, weak "parallel" correlations across contexts (e.g., 1a–2a, 1b–2b, 1c–2c) suggest that contextual differences in the stories may have altered how students responded even when the targeted indicator was nominally the same—an important diagnostic insight for item revision (Arini et al., 2023). In addition, the earlier-noted ceiling effect likely compressed correlations by reducing variance, making relationships statistically harder to detect (Carpenter et al., 2022). Taken together, the validity evidence supports use of the instrument, while the reliability, descriptive, and correlation patterns highlight clear improvement pathways: increase item difficulty gradient (especially for high-ceiling items), strengthen prompts/rubric requirements for evaluation and explanation, reduce context-driven ambiguity, and in future studies formally quantify agreement between AI and human scorers to ensure the scoring process remains both efficient and substantively accurate (Rivas et al., 2023).

CONCLUSION

Based on the findings and the analyses conducted to address the research questions, the study draws the following conclusions: the storytelling-based test instrument demonstrated strong content validity based on the Gregory index (0.80). Internal consistency reliability, assessed using Cronbach's Alpha, was α=0.623 (moderate/acceptable), indicating that the instrument was suitable for use in this study; however, further refinement—particularly of items with lower contribution—is recommended to strengthen internal consistency; students' critical thinking profiles showed the highest achievement in

Interpretation (93.6%), Inference (92.6%), and Analysis (92.2%), while performance was comparatively lower and more variable in Explanation (86.0%) and especially Evaluation (82.4%). Based on total-score categorisation (maximum score = 50), 88.37% of students were classified as very critical, 10.47% as critical, and 1.16% as moderately critical; inter-item relationships were generally weak according to Spearman correlation analysis, with most coefficients ranging from −0.12 to 0.30, although a limited number of item pairs showed stronger correlations above 0.30.

Suggestions for future improvements by researchers include aligning cross-context indicator pairs by standardizing the depth of requirements, the type of information to be used, and the form of evidence/concepts of electrochemistry that must appear in order to make measurements more stable across stories; revise items with low contribution (numbers 1d, 2d, and 2e) to consistency through a small pilot test after revision; the need for manual verification or prompt adjustments to make the assessment more discriminative, and the use of Gemini Pro instead of Gemini Flash to ensure more accurate and precise scoring.

REFERENCE

Aditya, R. (2021). Infrastruktur Cloud Pintar Dalam Sistem Layanan Informasi Berbasis Big Data. *Integr. J. Inf. Tech. Vocational. Educ.*, *3*(1), 29–38. https://doi.org/10.17509/integrated.v3i1.64423

Arini, R., Rahayu, Y. S., & Erman, E. (2023). Profile of Critical Thinking Results Analyzed From Facione Indicators and Gender of Learners. *Ijorer International Journal of Recent Educational Research*, *4*(4), 434–446. https://doi.org/10.46245/ijorer.v4i4.328

Azhar, K., Daud, M., Khidzir, N. Z., & Ismail, A. R. (2018). *Validity and reliability of instrument to measure social media skills among small and medium entrepreneurs at Pengkalan Datu River*. *7*(3), 1026–1037.

Beyera, G. K., O'Brien, J., & Campbell, S. (2020). The Development and Validation of a Measurement Instrument to Investigate Determinants of Health Care Utilisation for Low Back Pain in Ethiopia. *Plos One*, *15*(1), e0227801. https://doi.org/10.1371/journal.pone.0227801

Carpenter, C. M., Cooper, S. B., Wilson, P. L., Miller, S. M., Wyatt, C. W., Johnson, B. L., Shea, K. G., & Ellis, H. B. (2022). An Activity Scale for All Youth Athletes? Clinical Considerations for the HSS Pedi-Fabs. *Orthopaedic Journal of Sports Medicine*, *10*(12). https://doi.org/10.1177/23259671221143534

Casinillo, L. F. (2022). Calculus Teacher's Competencies as Correlates of Students' Learning Experiences. *International Journal of Indonesian Education and Teaching*, *7*(1), 22–32. https://doi.org/10.24071/ijiet.v7i1.5192

Chamberlain, D., Doeltgen, S., Knowles, R., Yaxley, A., & Miller, M. (2021). *Identification of Nutritional Risk in the Acute Care Setting: Progress Towards a Practice and Evidence Informed Systems Level Approach*. https://doi.org/10.21203/rs.3.rs-724006/v1

Creswell, J. W., & Plano Clark, V. L. (2023). Revisiting mixed methods research designs twenty years later. In *Handbook of mixed methods research designs* (Vol. 1, Issue 1).

Dewi, K. S., & Pramartha, I. N. B. (2021). *The Effectiveness of Authentic Project-Based Assessment on the Online Learning System Toward Learning Result of Student Viewed From Critical Thinking Ability*. https://doi.org/10.4108/eai.19-12-2020.2309123

Dolansky, M. A., Moore, S. M., Palmieri, P. A., & Singh, M. (2020). Development and Validation of the Systems Thinking Scale. *Journal of General Internal Medicine*, *35*(8), 2314–2320. https://doi.org/10.1007/s11606-020-05830-1

Eliza, E., Djudin, T., & Oktavianty, E. (2022). Development of Critical Thinking Ability Test Using the Rasch Model on Substance Pressure Materials. *Prisma Sains Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan Ipa Ikip Mataram*, *10*(4), 842. https://doi.org/10.33394/j-ps.v10i4.5987

Erwanto, E. (2020). Profil Kemampuan Berpikir Kritis Siswa Pada Konsep Keanekaragaman Hayati Melalui Problem Based Learning. *Jurnal Kependidikan Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan Pengajaran Dan Pembelajaran*, *6*(3), 578. https://doi.org/10.33394/jk.v6i3.2916

Gencer, A. S., & Doğan, H. (2020). The Assessment of the Fifth-Grade Students' Science Critical Thinking Skills Through Design-Based STEM Education. *International Journal of Assessment Tools in Education*, *7*(4), 690–714. https://doi.org/10.21449/ijate.744640

Gladly, A., Santo, N., & Octoria, D. (2024). *Analysis of High School Students ' Critical Thinking Skills on Capital Market Material*. *8*(2), 146–151.

Hanania, N., Nassir, M., Eghbarieh, N., & Masarwa, A. (2022). *A Stereodivergent Approach to the Synthesis of Gem-Diborylcyclopropanes*. https://doi.org/10.26434/chemrxiv-2022-cg7lm-v2

Huang, M. C., Chou, C., Wu, Y., Shih, J., Yeh, C. Y., Lao, A. C., Fong, H., Lin, Y.-F., & Chan, T. (2020). Interest-Driven Video Creation for Learning Mathematics. *Journal of Computers in Education*, *7*(3), 395–433. https://doi.org/10.1007/s40692-020-00161-w

Imanda, R., Setiawaty, S., & Qausar, H. (2024). Pendampingan Siswa Sekolah Dasar Melalui Pembelajaran Model Discovery Learning Berorientasi HOTS. *Jurnal Pengabdian Sosial*, *1*(4), 205–210. https://doi.org/10.59837/vqcmkj32

Kamalia, N., & Wasis, W. (2021). Analisis Keterampilan Berpikir Kritis Peserta Didik SMA Dalam Menyelesaikan Soal Fluida Statis. *Ipf Inovasi Pendidikan Fisika*, *10*(1), 90–98. https://doi.org/10.26740/ipf.v10n1.p90-98

Kristianawati, A., Priyayi, D. F., & Krave, A. S. (2022). Hubungan Keterampilan Laboratorium Dan Kemampuan Berpikir Kritis Siswa Pada Proses Pembelajaran Materi Jaringan Tumbuhan. *Didaktika Biologi Jurnal Penelitian Pendidikan Biologi*, *5*(2), 69. https://doi.org/10.32502/dikbio.v5i2.4389

Mutakin, D. (2023). Pemanfaatan Software SPSS Untuk Analisis Instrument Penelitian Dalam Penelitian Tindakan Kelas. *Jurnal Inovasi Dan Manajemen Pendidikan*, *3*(1), 33–49.

https://doi.org/10.12928/jimp.v3i1.7727

Nurhayati, I., Pramono, K. S. E., & Farida, A. (2024). Keterampilan 4C (Critical Thinking, Creativity, Communication and Collaboration) Dalam Pembelajaran IPS Untuk Menjawab Tantangan Abad 21. *Jurnal Basicedu*, *8*(1), 36–43. https://doi.org/10.31004/basicedu.v8i1.6842

Nuryuliana, N., & Prodjosantoso, A. K. (2021). Development of Integrated Assessment Instrument to Measure Critical Thinking Skills and Self-Efficiency in Acid-Base Concept. *JTK (Jurnal Tadris Kimiya)*, *6*(2), 223–235. https://doi.org/10.15575/jtk.v6i2.14872

Puspa, C. I. S., Rahayu, D. N. O., & Parhan, M. (2023). Transformasi Pendidikan Abad 21 Dalam Merealisasikan Sumber Daya Manusia Unggul Menuju Indonesia Emas 2045. *Jurnal Basicedu*, *7*(5), 3309–3321. https://doi.org/10.31004/basicedu.v7i5.5030

Radiansyah, R., Sari, R., Jannah, F., Prihandoko, Y., & Rahmaniah, N. F. (2024). Development of Evaluation Tools HOTS Based Project Learning Model to Improve Critical Thingking Ability. *Curriculum*, *3*(1), 1–8. https://doi.org/10.35335/curriculum.v3i1.127

Rahmawati, M., & Hidayati, Y. M. (2023). *Analysis of Mid-Semester Assessment in Cultivating Critical Thinking of Elementary School Students*. 1290–1299. https://doi.org/10.2991/978-2-38476-086-2_107

Rivas, S. F., Franco, A., Vieira, R. M., Almeida, L. S., & Saiz, C. (2023). Assessing Higher Education Students' Critical Thinking With the PENCRISAL Test - Portuguese Short Version: A Psychometric Study. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1196794

Rosidah, Sabri, S., & Ilmi, N. (2021). Description of Students' Critical Thinking Skills in Solving Open Ended Problems Based on Their Learning Style. *Arrus Journal of Mathematics and Applied Science*, *1*(2), 89–100. https://doi.org/10.35877/mathscience641

Rosli, M. S., Saleh, N. S., Alshammarı, S. H., Ibrahim, M. M., Atan, A. S., & Atan, N. A. (2021). Improving Questionnaire Reliability Using Construct Reliability for Researches in Educational Technology. *International Journal of Interactive Mobile Technologies (Ijim)*, *15*(04), 109. https://doi.org/10.3991/ijim.v15i04.20199

Salter, A., Fox, R. J., Tyry, T., Cutter, G., & Marrie, R. A. (2020). New Applications for Independent Activities of Daily Living in Measuring Disability in Multiple Sclerosis. *Multiple Sclerosis Journal*, *27*(1), 97–106. https://doi.org/10.1177/1352458519898591

Semegni, C. K. e., Phetlhu, D. R., & Modeste, R. R. M. (2021). An Integrative Review of Measurement Instruments Used to Assess the Stigma That Affects People Who Use Drugs. *Sage Open*, *11*(1). https://doi.org/10.1177/2158244020963067

Septiawati, L., & Eftanastarini, I. (2020). Analisis Ketercapaian Standar Kompetensi Lulusan Di MTS as Salam. *Attractive Innovative Education Journal*, *2*(1), 81. https://doi.org/10.51278/aj.v2i1.20

Sugiarta, I. M., Ariawan, I. P. W., Ardana, I. M., Divayana, D. G. H., Sukawijana, I. K. G., & Sugiharni, G. A. D. (2024). Validity and Reliability of the Discrepancy Evaluation Instrument for Measuring

Inequality in the Online Learning. *International Journal of Evaluation and Research in Education (Ijere)*, *13*(6), 3952. https://doi.org/10.11591/ijere.v13i6.28106

Suherman, S., Setiawan, R. H., Herdian, H., & Anggoro, B. S. (2022). 21st Century STEM Education: An Increase in Mathematical Critical Thinking Skills and Gender Through Technological Integration. *Journal of Advanced Sciences and Mathematics Education*, *1*(2), 33–40. https://doi.org/10.58524/jasme.v1i2.29

Sunarti, T., Amiruddin, M. Z. B., Wasis, W., Setyarsih, W., Jauhariyah, M. N. R., & Zainuddin, A. (2023). Discussion of Learning: Implementation and Effect on Critical Thinking Skills of Students in General Physics Courses. *Studies in Learning and Teaching*, *4*(2), 260–272. https://doi.org/10.46627/silet.v4i2.225

Susilawati, E., Agustinasari, A., Samsudin, A., & Siahaan, P. (2020). Analisis Tingkat Keterampilan Berpikir Kritis Siswa SMA. *Jurnal Pendidikan Fisika Dan Teknologi*, *6*(1), 11–16. https://doi.org/10.29303/jpft.v6i1.1453

Suyasa, P. W. A., & Divayana, D. G. H. (2022). Instrumen Evaluasi Model Discrepancy-Cse-Ucla Dalam Rangka Menunjang Evaluasi Efektivitas Pelaksanaan Pembelajaran Synchronous. *Jurnal Pedagogi Dan Pembelajaran*, *5*(2), 197–207. https://doi.org/10.23887/jp2.v5i2.48447

Szozda, A., Bruyere, K., Lee, H., Mahaffy, P. G., & Flynn, A. B. (2022). Investigating Educators' Perspectives Toward Systems Thinking in Chemistry Education From International Contexts. *Journal of Chemical Education*, *99*(7), 2474–2483. https://doi.org/10.1021/acs.jchemed.2c00138

Tleshova, Z., Tusselbayeva, Z., Ichshanova, A., Urazbekova, A., Zhenisbayeva, M., & Orymbayev, A. (2025). Reliability of Ai in Foreign Language Speaking Assessment: Comparing Automated and Human Scoring Among Undergraduate It Students in Kazakhstan. *National Center for Higher Education Development*, *2*(50). https://doi.org/10.59787/2413-5488-2025-50-2-18-31

Ulfa, N. J., & Permatasari, D. (2023). Critical Thinking Ability in Solving Word Problems Based on David Kolb's Learning Style. *Journal of General Education and Humanities*, *2*(4), 247–262. https://doi.org/10.58421/gehu.v2i4.70

Wahyudiati, D. (2022). Critical Thinking Skills and Scientific Attitudes of Pre-Service Chemistry Teachers Through the Implementation of Problem-Based Learning Model. *Jurnal Penelitian Pendidikan Ipa*, *8*(1), 216–221. https://doi.org/10.29303/jppipa.v8i1.1278

Wahyuni, D., Nurhuda, D. A., & Harini, H. (2022). PENGARUH FAKTOR INTERNAL TERHADAP KEMAHIRAN BERBAHASA INDONESIA PESERTA DIDIK PADA JENJANG SEKOLAH MENENGAH DI DKI JAKARTA (The Influence of Internal Factors Towards Indonesian Language Proficiency of Students at the High School in DKI Jakarta). *Sawerigading*, *28*(2). https://doi.org/10.26499/sawer.v28i2.1060

Zen, M. A., Wahyuningsih, S., & Dani, A. T. R. (2023). Aplikasi Pengelompokan Data Runtun Waktu

Dengan Algoritma K-Medoids. *Inferensi*, *6*(2), 117. https://doi.org/10.12962/j27213862.v6i2.15864

Zuhannisa', S., Jufriadi, A., & Budianto, H. (2023). Penerapan Model Pembelajaran Problem Based Learning (PBL) Berbantu Video Untuk Meningkatkan Kemampuan Berpikir Kritis Peserta Didik Pada Materi Pemanasan Global. *Jurnal Pembelajaran Bimbingan Dan Pengelolaan Pendidikan*, *3*(6), 473–483. https://doi.org/10.17977/um065v3i62023p473-483