International Journal of Pedagogy and Teacher Education

Journal Homepage: jurnal.uns.ac.id/ijpte

AI Scaffolding for Evidence-Based Critical Thinking in a Microcontroller Learning-Media Design Project

Harjito[1]*, Wahyu Hardyanto[2], Woro Sumarni[1], Sri Wardani[1]

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

[2]Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia

ARTICLE INFO

Article History

Received: February 08, 2026

1st Revision: March 02, 2026

Accepted: April 26, 2026

Available Online: April 30, 2026

Keywords:

Artificial intelligence in education; Critical thinking; STEAM learning; AI scaffolding; Project-based learning;

Corresponding Author

Email address:

harjito@mail.unnes.ac.id

ABSTRACT

The rapid integration of artificial intelligence (AI) into education has increased interest in how digital tools can support higher-order thinking rather than simply automate tasks. Within STEAM learning, this issue is especially important because students must combine evidence, reasoning, and design decisions in authentic project work. Generative AI tools based on large language models were used in this study as learning scaffolds to assist students with keyword expansion, scientific query construction, journal abstract summarization, hypothesis development, and revision of written explanations. The study examined whether such AI-supported learning improved students' critical thinking during a STEAM microcontroller media-design project. Two intact classes completed the same project workflow: one class used AI support, while the other completed the project without AI. Student performance was evaluated using a five-dimension critical-thinking rubric covering information and keywords, concepts and logic, journal abstract interpretation, hypothesis and reasoning, and academic formatting and compliance. Results showed that the AI-supported class performed better overall, scoring about five points higher on average and showing a large overall advantage. Improvements appeared across all rubric dimensions, with the strongest gains found in journal abstract interpretation and hypothesis and reasoning, suggesting that AI was most helpful when students had to interpret evidence, connect ideas, and justify decisions. The weakest-link bottleneck analysis, defined as the rubric dimension in which each student performed worst, showed that journal abstract interpretation remained the main constraint in both groups. These findings indicate that AI can serve as a productive scaffold for critical thinking and decision-making in STEAM projects, especially by supporting evidence use and iterative reasoning, while also highlighting the need for explicit instruction in reading and interpreting scientific abstracts.

How to cite: Harjito, Hardyanto, W., Sumarni, W., & Wardani, S. (2026). AI Scaffolding for Evidence-Based Critical Thinking in a Microcontroller Learning-Media Design Project. *International Journal of Pedagogy and Teacher Education, 10*(1), 88–101. https://doi.org/10.20961/ijpte.v10i1.115698

1. INTRODUCTION

Critical thinking has become a central outcome in science and STEAM education because students are increasingly expected to evaluate information, justify decisions, and revise ideas in response to evidence. Educational scholarship consistently defines critical thinking as purposeful and reasoned judgment involving interpretation, analysis, evaluation, inference, and explanation (Facione, 1990; Ennis, 1993). Psychological perspectives further emphasize that critical thinking depends not merely on intelligence, but on the deliberate use of strategies for problem solving and decision making under uncertainty (Halpern, 2014). Classroom assessment therefore needs to capture the quality of students' reasoning rather than only the correctness of final products. Performance-based tasks and analytic rubrics are particularly suitable for this purpose because they make students' evidence use, justification, and decision-making processes visible (Brookhart, 2010; Paul & Elder, 2006).

STEAM learning offers a productive context for developing such thinking because it requires students to integrate scientific concepts, engineering constraints, technological tools, and communication within authentic problem-solving situations. A microcontroller-based media-design project is especially relevant because students must not only create a functional product, but also search for information, interpret scientific evidence, justify design choices, and explain how those choices are supported by disciplinary knowledge. Project-based and integrated STEAM approaches have long argued that deeper learning emerges when students work on meaningful tasks, produce artifacts, and receive feedback on both design and reasoning (Thomas, 2000; Honey

p-ISSN: 2597-7792 / e-ISSN: 2549-8525

DOI: https://doi.org/10.20961/ijpte.v10i1.115698

et al., 2014). Science education frameworks likewise position explanation, argument from evidence, and modeling as core practices for meaningful learning (National Research Council, 2012). Authentic STEAM work, however, does not automatically produce strong reasoning. Maker- and construction-oriented environments can stimulate creativity and iterative thinking, yet they may also drift toward procedural activity unless learners are explicitly supported in connecting evidence, concepts, and decisions (Papert, 1980; Blikstein, 2013).

Artificial intelligence introduces a new layer of support within this context, especially through generative AI and reasoning-assistance tools that can assist students during inquiry and design processes. In STEAM classrooms, such tools can be used to expand keywords, formulate search queries, summarize journal abstracts, compare claims across sources, generate alternative explanations, and revise written justifications. When used in this way, AI functions less as an answer machine and more as a scaffold for reasoning. Prior work in AI in education suggests that AI may act as a tutor, tool, or partner, depending on how it is embedded in pedagogy and how much cognitive responsibility remains with learners (Holmes et al., 2019). Large language models are particularly promising because they can accelerate iteration and support explanation building, but they also introduce serious concerns, including over-reliance, superficial acceptance of outputs, privacy risks, and unequal access (Kasneci et al., 2023). Responsible implementation therefore requires clear mitigation strategies, including teacher supervision, source verification, explicit evidence-checking routines, and task designs that preserve learner agency and accountability. These expectations align closely with UNESCO guidance on transparency, fairness, human oversight, and the responsible use of generative AI in educational settings (UNESCO, 2021, 2023).

Toulmin's model offers a particularly relevant framework for examining whether AI support genuinely improves critical thinking in STEAM learning. Strong reasoning depends on the ability to connect claims to evidence through defensible justification, rather than merely producing fluent or persuasive language (Toulmin, 1958). Science education research similarly shows that students' argumentation improves when instruction explicitly supports the construction, critique, and refinement of evidence-based claims (Osborne et al., 2004). Explanation-focused approaches further demonstrate that learners benefit when the structure of claim, evidence, and reasoning is made explicit through scaffolds and assessment criteria (McNeill & Krajcik, 2008). This issue becomes especially important when students engage with scientific abstracts, where they must identify purpose, method, findings, and implications before that information can meaningfully inform a design decision. Generative AI may help students process such texts more efficiently, but polished output does not necessarily indicate stronger reasoning. A credible evaluation of AI support must therefore focus on whether students demonstrate better evidence selection, stronger justification, and more defensible decisions, rather than simply better wording or faster completion. Estimation-focused and robust analytic reporting is particularly appropriate for this purpose because it supports careful interpretation of learning gains without overstating results through single-threshold significance claims (Wasserstein & Lazar, 2016).

Current discussions on AI in education still tend to emphasize general promise or broad concerns, while offering less evidence about how AI operates within authentic production-oriented STEAM tasks and which dimensions of critical thinking it strengthens most. Limited attention has also been given to whether AI support changes not only overall performance, but also the structure of students' reasoning difficulties. This study addresses that gap by examining AI-supported and non-AI learning in a STEAM microcontroller media-design project, with critical thinking assessed through a multidimensional rubric focused on evidence use, conceptual logic, abstract interpretation, hypothesis construction, and academic compliance. The study makes three contributions. First, it clarifies how generative AI can be integrated into STEAM learning as a scaffold for evidence gathering, abstract interpretation, hypothesis development, and revision of explanations. Second, it identifies which components of critical thinking are most responsive to AI support and which bottlenecks remain persistent. Third, it offers a practically relevant model for teachers by showing that AI is most effective when embedded within structured routines that require students to verify sources, connect evidence to claims, justify design decisions, and revise their work through guided feedback. Such an approach positions AI as a support for disciplined reasoning rather than a substitute for it.

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

2. MATERIAL AND METHOD

Research Design

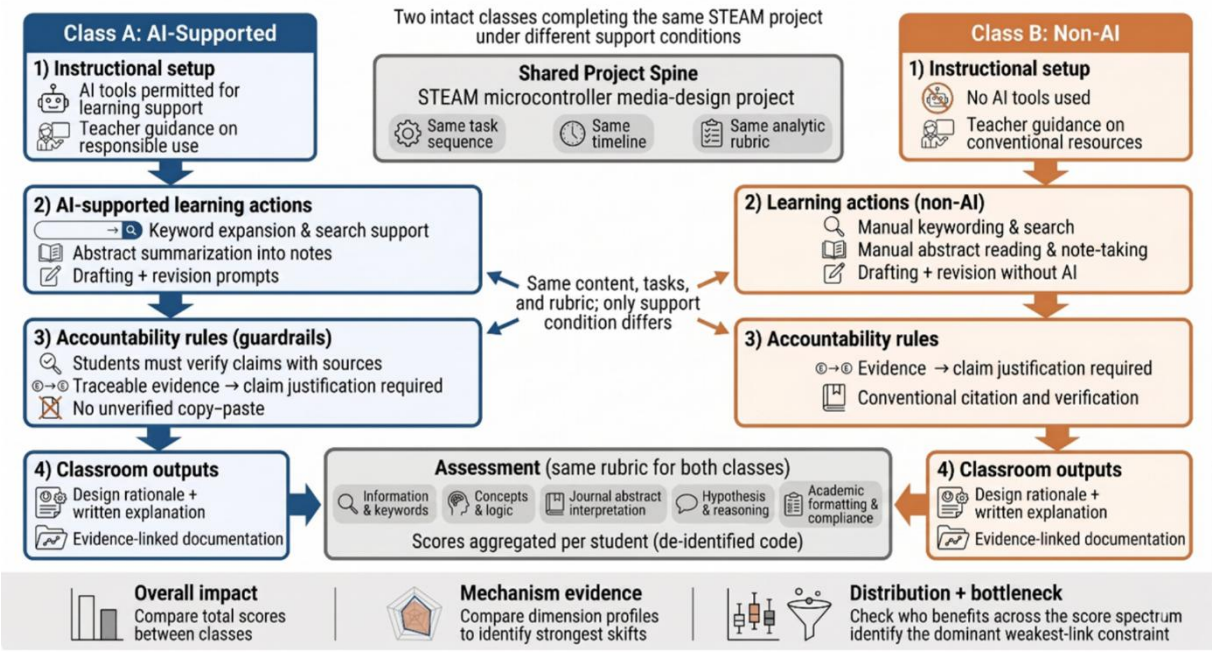

Figure 1. Research Design of the AI-Supported vs Non-AI STEAM Microcontroller Project Study

This study employed a comparative, non-randomized design using two intact classes, which may introduce selection-related bias. Several steps were therefore taken to minimize this risk. Both classes were taught within the same instructional setting, completed the same STEAM microcontroller media-design project, followed the same workflow, and were subject to the same project brief, timeline, deliverables, and analytic scoring rubric. The only planned difference between conditions was the availability of AI as a structured scaffold in Class A, whereas Class B completed the same task without generative AI support. In addition, scoring focused on the quality of reasoning demonstrated in the final documentation and artifacts rather than on surface features of language production. These design choices strengthen procedural comparability across groups, although they do not eliminate all bias associated with the intact-class design. Accordingly, the study is interpreted as providing robust comparative evidence under shared instructional constraints rather than definitive causal inference.

The study was conducted under two instructional conditions within the same project and analysis framework, as illustrated in Figure 1. Class A (AI-supported) used generative AI tools, specifically large language model–based chat assistants, as structured scaffolds for keyword expansion, search-query construction, abstract summarization, hypothesis framing, and revision of written explanations. All AI-assisted outputs had to be verified against source materials, and students remained responsible for articulating and justifying their final decisions in their own words. Class B (Non-AI) completed the same project without generative AI, relying instead on teacher guidance, textbooks or handouts, peer discussion, and manual web searches. Procedural comparability was maintained by ensuring that both classes worked with the same project brief, timeline, deliverables, and scoring rubric, so that the principal planned difference between conditions was the availability of AI support. The primary outcome was evidence-based critical thinking, assessed through a validated analytic rubric measuring students' ability to identify and interpret relevant information, apply concepts logically, formulate hypotheses, and justify design decisions in an academically responsible manner. For example, strong performance in Journal Abstract Interpretation required students in both conditions to accurately identify a study's purpose, method, and findings and relate them explicitly to the project decision, whereas strong performance in Hypothesis and Reasoning required a defensible claim supported by relevant evidence and coherent justification.

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

Participants and Data Sources

The unit of analysis in this study was the individual student. The final analytic sample consisted of students drawn from two intact classes participating in the same STEAM microcontroller media-design project. After data screening, the sample included 30 students in Class A (AI-supported) and 27 students in Class B (Non-AI). To ensure confidentiality and ethical compliance, all participants were represented using de-identified alphanumeric codes (e.g., M2301), and no personally identifiable information was used in the analysis. The unequal group sizes reflect natural classroom enrollment and were addressed analytically through estimation-based and distribution-focused methods rather than assumptions of equal sample sizes.

Two primary data sources were used for analysis. First, a categorical variable indicating group membership (AI-supported vs. Non-AI) was used to define the instructional condition. Second, students' rubric-based critical thinking scores served as the outcome data, including both the total rubric score (0–100) and the scores for each named rubric dimension. Students were included in the analytic dataset only if they had a complete rubric assessment and a valid group label. Records with a missing total rubric score were excluded from primary analyses to preserve the integrity and comparability of outcome estimates. This screening process resulted in a clean, complete dataset suitable for robust data science and machine-learning–oriented analyses of group differences in critical thinking performance.

Intervention and Procedure (AI vs. Non-AI Workflow)

Both instructional conditions followed the same project workflow, designed to elicit evidence-based critical thinking through an authentic STEAM task. Students progressed through a structured sequence of activities: project planning, evidence gathering, interpretation of information, formulation of a hypothesis or design rationale, and final product development accompanied by written documentation. This workflow ensured that all students engaged with the same cognitive demands, including identifying relevant information, making sense of evidence, and justifying design decisions. The final submission in both groups consisted of a learning-media artifact supported by a coherent written explanation of the underlying reasoning and design choices.

The intervention differed only in the availability of AI support. In Class A (AI-supported), AI tools were explicitly positioned as scaffolding mechanisms rather than as sources of final answers. Students were permitted to use AI for keyword expansion and scientific query construction, structured summarization of journal abstracts into evidence notes, hypothesis framing using claim–evidence–reasoning logic, and iterative revision of explanations. Crucially, students were required to verify, edit, and justify all AI-assisted outputs, and rubric scoring targeted the quality of reasoning rather than surface features of the text. In contrast, Class B (Non-AI) completed the same project using teacher guidance, textbooks or handouts, peer discussion, and manual note-taking, supplemented by conventional web searches without generative AI assistance. To maintain equivalence across conditions, both groups used the same task requirements, timeline, and rubric criteria, and all rubric scoring was conducted after final submission based solely on the completed written documentation and project artifacts.

Data Analysis Techniques

The primary outcome of this study was students' total critical-thinking performance, operationalized as a rubric score ranging from 0 to 100. The instrument was an analytic rubric designed to capture evidence-based reasoning quality in the written documentation accompanying the STEAM microcontroller media-design project. Rather than using generic or numbered "aspects," the rubric was organized into five interpretable dimensions aligned with Paul–Elder principles of disciplined reasoning (e.g., clarity, relevance, logic, depth, and justification). Each dimension had a defined maximum score, and the total rubric score was calculated as the sum of the five dimension scores, thereby producing a single composite indicator while preserving diagnostic information about where critical thinking was strongest or weakest.

To support transparency and replicability, Table 1 summarizes the rubric structure, maximum scores, and high-level indicators of strong performance. In scoring, the rater evaluated the extent to which students demonstrated purposeful information selection, conceptually accurate reasoning, disciplined interpretation of evidence, defensible hypothesis construction, and academically responsible presentation. Scoring consistency was maintained through prior calibration using anchor samples representing low-, medium-, and high-quality responses, which were used to stabilize interpretation of the rubric criteria before full scoring. To reduce

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

subjective inflation, a random 20–30% subset of responses was re-scored after an interval, and the consistency of these scores was checked to confirm internal scoring stability.

Table 1. Critical thinking rubric used for scoring project documentation (Paul–Elder–aligned)

Dimension (what it measures)	Max score	Indicators of strong performance (summary)
Information & keywords	20	Selects accurate, relevant keywords from multiple sources; maintains cross-source consistency; organizes evidence clearly in a structured table.
Concepts & logic	20	Uses correct scientific concepts; builds coherent logical connections; constructs effective scientific queries (AND/OR) aligned with the analysis goal.
Journal abstract interpretation	25	Accurately captures purpose–method–results from abstracts; synthesizes 1–3 studies objectively; connects evidence to the project's scientific focus.
Hypothesis & reasoning	25	Produces defensible hypotheses/claims grounded in evidence; provides rationale, implications, and alternatives; maintains internal logic consistent with critical-thinking standards.
Academic formatting & compliance	10	Uses clear, consistent academic formatting (e.g., DOI, paragraphing, alignment); follows instructions; presents a clean and readable structure.
Total	**100**	Sum of all dimensions.

Data Processing and Analysis Plan (Data Science + Robust Inference)

All analyses followed a reproducible data-science workflow designed to ensure transparency, robustness, and sensitivity to distributional features of the data. Raw rubric records were merged using de-identified student codes so that each observation corresponded to a single student. During preprocessing, group labels (AI-supported and non-AI) were verified, and all dimension-level and total scores were checked to confirm compliance with the defined rubric ranges. When duplicate entries were identified for a given student code, the record retained for analysis was the one exhibiting the highest internal consistency and completeness across rubric dimensions. Students lacking a total critical-thinking score were excluded from primary outcome analyses to avoid ambiguity in estimating group differences. The principal estimand was the difference in mean total critical-thinking scores between groups, expressed as $\Delta_{mean} = \bar{X}_{AI} - \bar{X}_{NonAI}$. Statistical uncertainty was characterized using a nonparametric bootstrap 95% confidence interval, while inferential robustness was assessed through a permutation test that does not rely on distributional normality assumptions. Effect magnitude was summarized with Hedges' g, providing a standardized estimate appropriate for finite samples and potential group-size imbalance.

Beyond mean-level comparisons, distribution-level evidence was examined to capture heterogeneity in outcomes. Empirical cumulative distribution functions (ECDFs) were compared between groups, with the Kolmogorov–Smirnov statistic serving as a global indicator of distributional separation. To examine whether AI support benefited students similarly across different performance levels, quantile treatment effects were estimated across quantiles q = 0.1–0.9 with bootstrap confidence bands. In this study, quantile treatment effects indicate how large the AI-related score difference was among lower-, middle-, and higher-performing students, rather than only at the overall mean. This approach was used to determine whether the effect of AI was broadly distributed or concentrated in particular parts of the score distribution. Component-level mechanisms were explored by estimating dimension-specific mean differences with confidence intervals and synthesizing these estimates in forest and profile visualizations. In addition, a weakest-link bottleneck analysis identified each student's lowest normalized rubric dimension and summarized the prevalence and severity of bottlenecks, yielding an interpretable fingerprint of persistent constraints on critical-thinking performance. Results were communicated through diagnostic visualizations, including estimation plots, ECDF comparisons, quantile-effect curves, dimension-level profiles, and the bottleneck fingerprint figure.

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

3. RESULTS

Overall Impact on Total Critical-Thinking Performance

The primary analysis was designed to estimate the magnitude and robustness of the AI effect on students' total critical-thinking performance, prioritizing effect size and uncertainty over binary "significant/non-significant" conclusions (Wasserstein & Lazar, 2016; Lakens et al., 2018). This estimation-first framing is particularly appropriate for classroom comparisons with moderate sample sizes, where practical significance and distributional behavior matter for interpretation (Lakens, 2013). Accordingly, we analyzed the total rubric score (0–100) as the main outcome to represent students' evidence-based critical thinking demonstrated in the STEAM microcontroller media-design project. The key result is summarized visually in Figure 2, which integrates raw-score distribution with the estimated mean difference and its uncertainty interval.

Students in the AI-supported class achieved higher total critical-thinking scores than those in the Non-AI class, with a mean advantage of 5.08 points (95% bootstrap CI [2.41, 7.87]). This difference corresponded to a large effect size (Hedges' g = 0.94), indicating a substantively meaningful benefit of AI support. The pattern suggests that AI functioned as a scaffold for evidence-based reasoning by supporting information processing and iterative refinement during the STEAM project (Paul & Elder, 2006; Halpern, 2014). This interpretation is also consistent with the view that appropriate scaffolding can reduce extraneous cognitive load and help students allocate more cognitive resources to higher-order reasoning (Sweller et al., 2011).

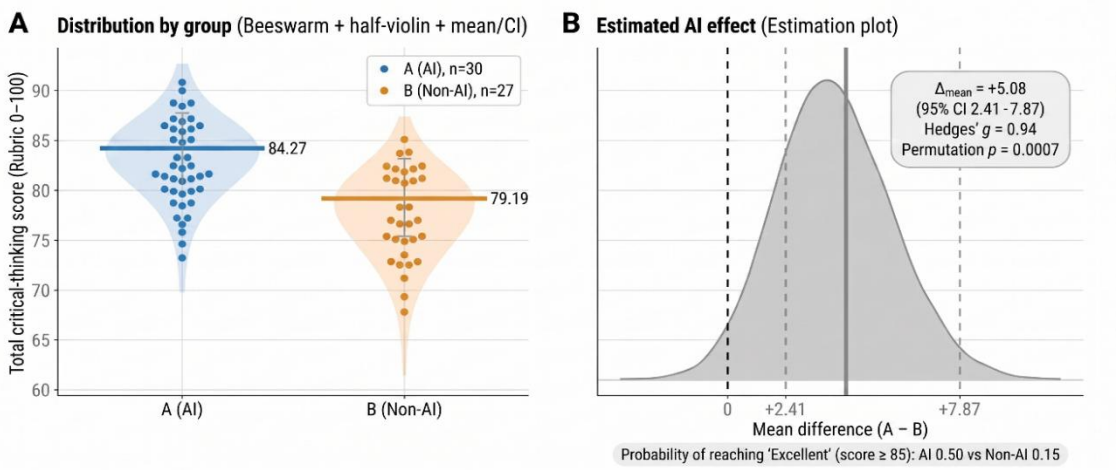

Figure 2. Estimation Plot of Total Critical-Thinking Scores

Critical-Thinking Characteristics Profiling (Dimension-Level Mechanism Evidence)

This analysis examined which dimensions of critical thinking were most strongly associated with AI support, beyond the overall total score. Dimension-level scores were normalized to their maximum values and summarized as mean percentages with 95% bootstrap confidence intervals (Lakens, 2013; Wasserstein & Lazar, 2016). The dimension-level pattern is visualized in Figure 3, showing the normalized critical-thinking profile of the AI-supported and non-AI classes with 95% bootstrap confidence intervals. The most important finding was that the AI-supported class showed its largest advantages in the evidence-intensive dimensions, particularly Journal Abstract Interpretation and Hypothesis and Reasoning. These dimensions require students to interpret scientific information, connect evidence to claims, and justify design decisions, making them especially relevant to the study's focus on evidence-based critical thinking (McNeill & Krajcik, 2008). This pattern suggests that AI support was most beneficial not for surface-level performance, but for the reasoning components most closely tied to evidence use and justification.

The resulting profiles indicate a consistent upward shift across all five dimensions for the AI-supported class relative to the non-AI class, suggesting broad-based improvement rather than gains confined to a single aspect of performance. This pattern rules out an interpretation in which AI primarily enhances surface features or procedural compliance, and instead points to a generalized strengthening of critical-thinking performance across components. The largest separations between groups are concentrated in the most evidence-intensive

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

dimensions: Journal abstract interpretation (Δ ≈ +8.3 percentage points) and Hypothesis & reasoning (Δ ≈ +5.1 percentage points). These dimensions require students to extract meaning from scientific sources, integrate evidence, and construct defensible claims, all of which impose substantial cognitive demands. The concentration of gains in these components is consistent with cognitive load accounts in which scaffolding reduces extraneous demands and frees cognitive resources for higher-order reasoning (Sweller, 1988). It also accords with recent empirical discussions of large language models as tools that can support drafting, revising, and synthesizing information when students remain responsible for evaluative judgment and justification (Kasneci et al., 2023). Together, the dimension-level profile provides mechanism-relevant evidence that AI support primarily strengthened students' capacity for evidence-based reasoning rather than merely improving formal or presentational aspects of their work.

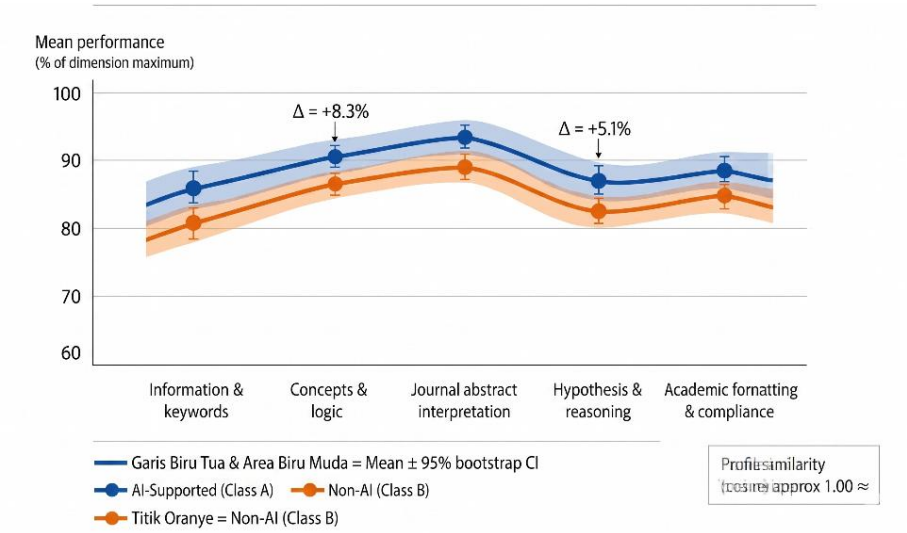

Figure 3. Normalized Critical Thinking Profile by Dimension with 95% Bootstrap CI

AI Effects Across The Entire Performance Distribution (Quantile Treatment Effect)

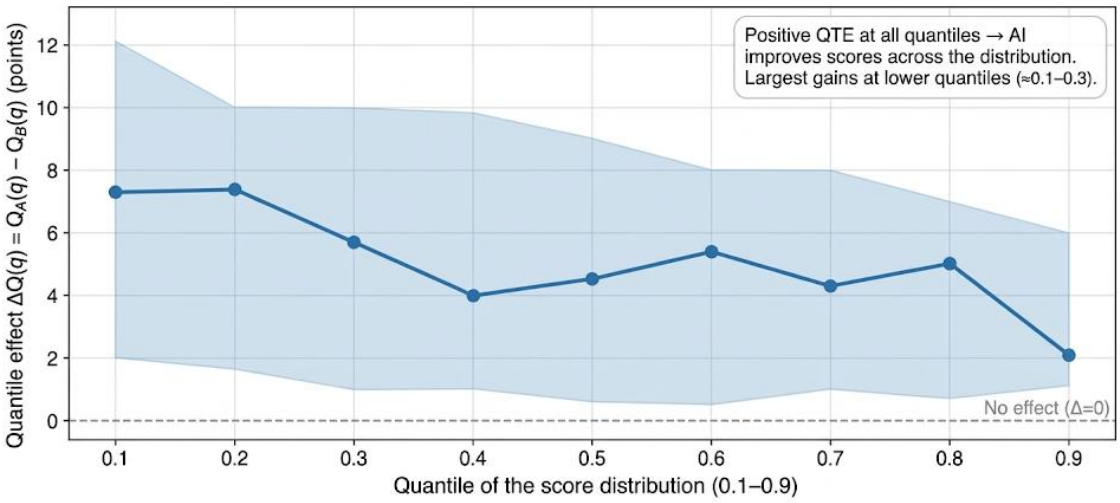

Figure 4. Quantile Treatment Effects of AI on Total Critical-Thinking Scores

This analysis examined whether the effect of AI support was confined to average performance or extended across the full achievement distribution. In educational settings, interventions frequently produce

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

heterogeneous effects, and reliance on mean differences alone can obscure where improvements actually occur. A distribution-aware approach was therefore adopted by estimating a Quantile Treatment Effect (QTE) curve, defined at each quantile level qas $\Delta Q(q) = Q_{AI}(q) - Q_{NonAI}(q)$. Quantile-based comparisons are particularly appropriate for instructional research because they reveal differential effects among lower-, middle-, and higher-performing students that are not visible in mean-based summaries (Koenker & Bassett, 1978; Firpo, 2007). Statistical uncertainty was characterized using bootstrap resampling, which provides robust confidence intervals without strong parametric assumptions (Efron, 1979). This estimation-oriented strategy is consistent with contemporary reporting guidance that emphasizes effect patterns with uncertainty rather than reliance on a single null-hypothesis decision rule (Wasserstein & Lazar, 2016).

The QTE results shown in Figure 4 indicate positive effects across all examined quantiles (0.1–0.9), showing that AI support was associated with improved critical-thinking performance across the score distribution. In educational research, QTE is useful because it shows whether an intervention works similarly for lower-, middle-, and higher-performing students, rather than only at the average level. In this study, the largest gains appeared at the lower quantiles (approximately q = 0.1–0.3), suggesting that AI support provided the strongest scaffolding for students with weaker initial performance. Positive, though smaller, effects were also observed toward the upper quantiles, indicating that the benefit of AI was not limited to one subgroup. Taken together, this pattern suggests that AI support functioned as a broadly beneficial scaffold across performance levels, with the strongest added value for lower-to-mid performers. This distributional pattern complements the mean-difference results by showing that the observed benefit was not only an average effect, but one that extended across different levels of student achievement (Chernozhukov et al., 2013).

Dimension-Level AI Effects On Critical-Thinking Components (Mechanism Evidence)

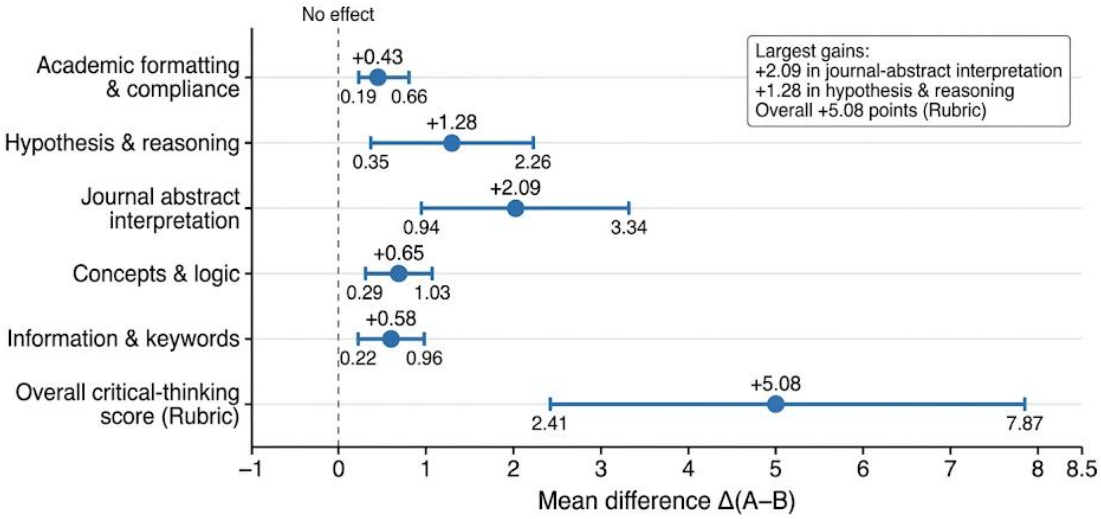

Figure 5. AI Effects on Critical Thinking Dimensions with 95% Bootstrap CIs

This analysis was conducted to clarify which specific components of critical thinking were most strengthened under AI support and to support a mechanism-oriented interpretation of the overall group difference. Because critical thinking is inherently multidimensional, changes in a composite total score may conceal which reasoning processes are actually shifting. Dimension-level decomposition therefore provides a more diagnostically useful account of intervention impact, particularly for STEAM tasks in which evidence use and explanation quality are central learning outcomes (McNeill & Krajcik, 2008). Estimation was framed in terms of effect magnitudes accompanied by uncertainty intervals, consistent with methodological guidance that prioritizes transparent estimation over dichotomous "significant/non-significant" conclusions (Lakens, 2013; Wasserstein & Lazar, 2016).

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

Dimension-level effects were estimated as mean differences (AI − Non-AI) for each rubric component, and uncertainty was quantified using bootstrap confidence intervals. The pattern in Figure 5 shows positive differences across all dimensions, indicating broad improvement rather than gains limited to surface presentation or procedural compliance. The largest effects occur in the most evidence-intensive components—journal abstract interpretation (Δ = +2.09, 95% CI 0.94 to 3.34) and hypothesis & reasoning (Δ = +1.28, 95% CI 0.35 to 2.26)—which aligns with an interpretation that AI primarily supported evidence extraction, synthesis, and argument construction. More moderate but consistently positive gains were observed for concepts & logic (Δ = +0.65, 95% CI 0.29 to 1.03) and information & keywords (Δ = +0.58, 95% CI 0.22 to 0.96), while academic formatting & compliance showed the smallest shift (Δ = +0.43, 95% CI 0.19 to 0.66).

Taken together, the component profile supports a theoretically coherent account of how AI support functioned in this learning context. The concentration of larger gains in evidence-intensive reasoning suggests that AI contributed most when students faced high cognitive demands associated with interpreting scientific sources and maintaining coherent justification under limited working-memory resources (Sweller, 1988). At the same time, the results are consistent with contemporary analyses of large language models in education that characterize their strongest pedagogical value as facilitating iterative explanation-building and feedback-like refinement, provided that student accountability for reasoning is preserved (Kasneci et al., 2023).

Bottleneck Funnel Diagnosis: Weakest-Link Constraints in Critical-Thinking Performance

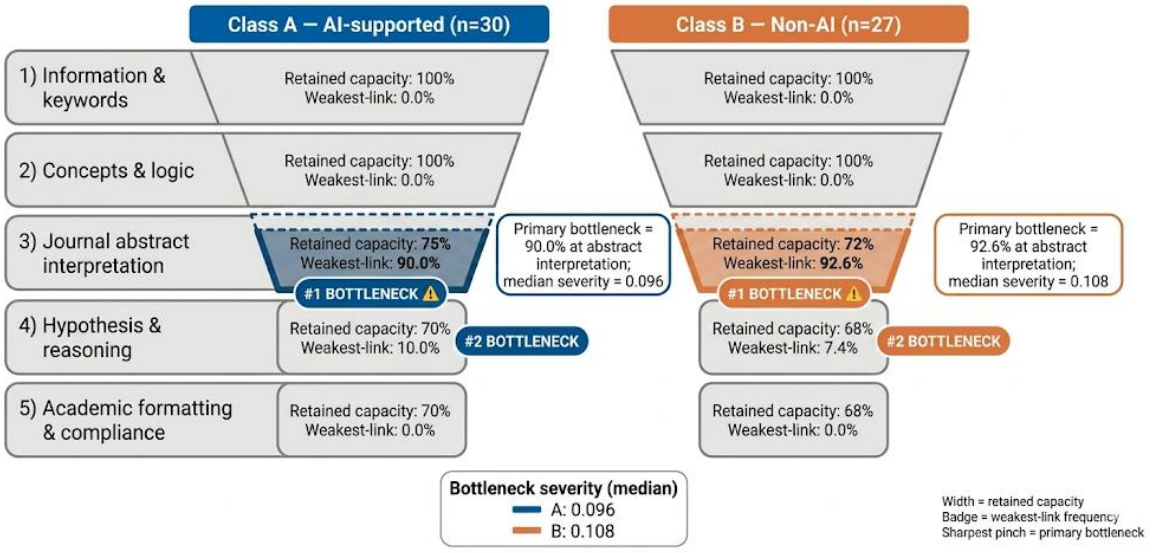

Figure 6. Critical Thinking Bottleneck Funnel: AI vs Non-AI Groups

Bottleneck analysis revealed a consistent pattern across both conditions: Journal Abstract Interpretation remained the dominant weakest dimension. This suggests that students' ability to identify and interpret evidence from scientific abstracts continued to limit their broader critical-thinking performance. Although AI support lessened the intensity of this constraint, it did not remove it. Targeted instructional support in abstract reading and evidence extraction therefore remains necessary (Sweller, 1988; McNeill & Krajcik, 2008).

To operationalize this objective, a bottleneck funnel analysis was conducted by identifying, for each student, the rubric dimension with the lowest normalized score and summarizing the population-level prevalence of these weakest dimensions for both instructional conditions. Figure 6 visualizes the resulting bottleneck fingerprints. The analysis reveals a highly concentrated constraint in Journal abstract interpretation for both classes: this dimension constituted the weakest link for 90.0% of students in the AI-supported class and 92.6% of students in the non-AI class. Hypothesis & reasoning emerged only as a secondary bottleneck, affecting a much smaller proportion of students (10.0% in the AI-supported class and 7.4% in the non-AI class). Consistent with these prevalence patterns, bottleneck severity—summarized by the median weakest-link gap—was

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

modestly lower in the AI-supported class (0.096) than in the non-AI class (0.108). Together, these results indicate that AI support reduces the intensity of the dominant constraint but does not alter its identity.

The bottleneck analysis complements the overall findings by showing that improved performance under AI support did not eliminate the core difficulty in Journal Abstract Interpretation. Although AI appeared to assist students in keyword generation, drafting, and revision, it did not fully address the more demanding task of extracting and synthesizing evidence from scientific abstracts. This pattern suggests that AI can support parts of the reasoning process, but explicit instructional guidance is still needed to strengthen abstract interpretation as a central component of critical thinking (Efron, 1979; Lakens, 2013).

4. DISCUSSION

Overall Impact of AI Support on Critical-Thinking Performance

The primary outcome indicates that students in the AI-supported class achieved higher total critical-thinking performance than their non-AI peers. In the total rubric score (0–100), the estimated mean difference was Δmean (AI − Non-AI) = +5.08 points, with a 95% bootstrap confidence interval of [2.41, 7.87], and this result was further supported by a permutation test ($p = 0.0007$), as illustrated in Figure 2. An estimation-first interpretation places greater emphasis on effect magnitude and uncertainty than on dichotomous thresholding alone, in line with current recommendations to interpret educational effects through confidence intervals and practical meaning (Cumming, 2014). From an effect-size perspective, the standardized difference was Hedges' $g = 0.94$, which is conventionally interpreted as a large effect when comparing group means under moderate sample sizes, particularly when applying small-sample bias correction (Hedges, 1981). This pattern strengthens the conclusion that the observed gap was not a trivial statistical artifact, but reflected a meaningful improvement in students' demonstrated critical-thinking quality under the project conditions.

A defensible learning-science interpretation is that AI acted primarily as an iterative feedback and self-regulation scaffold, reducing friction in drafting, checking, and revising so that students could allocate more effort to evidence-based reasoning and refinement. In this project, students used AI in concrete ways, such as expanding search keywords, generating alternative search queries, summarizing journal abstracts into provisional evidence notes, checking whether a drafted explanation was consistent with the cited source, and revising hypotheses to make the connection between evidence and design decisions more explicit. These uses suggest that AI did not replace students' reasoning, but instead supported intermediate steps in the reasoning process. This interpretation aligns with feedback theory, which emphasizes that learning gains depend on feedback directed at the task, process, and self-regulation levels (Hattie & Timperley, 2007), and with self-regulated learning models in which feedback supports monitoring and strategic control during complex tasks (Butler & Winne, 1995). In classroom terms, AI appeared most useful when it accelerated cycles of attempt, feedback, and revision, thereby functioning similarly to formative feedback structures that support student agency and self-regulation, a pattern that is consistent with the score separation presented in Figure 2 (Nicol & Macfarlane-Dick, 2006). This advantage, however, remained pedagogy-dependent, because the educational value of generative AI depends on task design, verification routines, and accountability structures rather than on AI access alone (Albadarin et al., 2024; Dignath et al., 2008).

Mechanisms of Improvement at the Critical-Thinking Dimension Level

The dimension-level pattern suggests a general uplift across rubric components, with the most pronounced gains concentrated in evidence-intensive dimensions—those requiring students to extract meaning from scientific texts and convert that information into defensible reasoning. This upward shift appears consistently across all five dimensions in the normalized profile, indicating that the advantage of AI support was not confined to a single aspect of performance, as reflected in Figure 3. Such a pattern is theoretically coherent because these dimensions demand coordinated epistemic work, including selecting relevant information, evaluating what counts as evidence, and connecting evidence to warrants and claims, which is precisely where scaffolding and iterative feedback are most likely to improve performance quality rather than merely strengthen surface compliance (Berland & Reiser, 2009; Sandoval & Millwood, 2005).

A more diagnostic reading of the results indicates that AI-related gains were strongest in the dimensions requiring students to engage directly with evidence. Students appeared to improve particularly in reading scientific abstracts, identifying key elements such as purpose, method, and results, and using that information

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

to support explanations and design decisions. This interpretation is reinforced by the dimension-specific mean-difference estimates presented in Figure 5, where the largest gains are concentrated in Journal Abstract Interpretation and Hypothesis & Reasoning. Such a pattern is closely aligned with the core features of critical thinking, because high-quality reasoning requires students not only to formulate claims, but also to justify those claims with relevant, traceable, and conceptually coherent evidence (Berland & Reiser, 2009; Cavagnetto, 2010). Research on evidence use similarly shows that students' reasoning improves when instructional support is directed toward selecting, interpreting, and mobilizing evidence in written explanations (Sandoval & Millwood, 2005). Read together, the profile pattern in Figure 3 and the component-specific effect estimates in Figure 5 suggest that AI support was most beneficial not for surface-level presentation, but for the reasoning processes most closely tied to evidence use and justification. This result carries an important implication for assessment in AI-supported STEAM contexts: evaluation should prioritize evidence use, coherence of justification, and explicit links between source text and inference, rather than giving disproportionate weight to formal presentation alone (Panadero & Jonsson, 2013).

Systemic Bottleneck in Journal-Abstract Interpretation and Targeted Intervention Design

The sequence of problem-based learning includes problem identification, brainstorming, surfacing prior The benefits of AI support were not confined to the average score difference, but extended across the performance distribution. Positive quantile treatment effects were observed throughout the examined quantiles, with the largest gains concentrated among lower-to-mid performers, indicating that AI provided its greatest added value for students who required more support during complex reasoning tasks, as shown in Figure 4. This pattern strengthens the practical significance of the findings because it suggests that AI functioned as a broadly accessible scaffold rather than as a tool that benefited only already high-performing students. At the same time, the uneven magnitude of the gains across quantiles implies that AI support did not operate uniformly, but was especially helpful when students were struggling with demanding intermediate processes such as locating relevant evidence, organizing information, and refining tentative explanations.

A more constrained picture emerges, however, when the results are examined through the bottleneck fingerprint. Journal Abstract Interpretation remained the dominant weakest link in both groups, accounting for the limiting dimension for the vast majority of students, while other components contributed comparatively little to being the principal constraint, as reflected in Figure 6. This pattern is theoretically plausible in project-based STEAM tasks because abstract interpretation imposes high intrinsic cognitive load: students must process dense information, unfamiliar disciplinary discourse, and compressed accounts of methods and findings while simultaneously deciding what is relevant for their own reasoning (Sweller, 1988; van Merriënboer & Sweller, 2005). When cognitive load remains high and students do not yet possess stable schemas for parsing scientific texts, cognitive resources are diverted away from deeper integration of evidence, warrants, and claims, allowing abstract interpretation to persist as the central bottleneck even when other supports are available.

The persistence of this bottleneck suggests that the limitations of AI support are best addressed not by reducing its use, but by embedding it within more structured instructional routines. Although AI appeared to assist students in keyword development, drafting, and information organization, it was less effective in helping them interpret scientific abstracts and use them as evidence for justified design decisions. A more productive pedagogical response is therefore to combine AI with explicit routines for abstract reading, such as identifying the purpose, method, results, and claim of a study, linking source evidence to student claims, and checking whether a conclusion is adequately supported. The contrast between the broad positive reach shown in Figure 4 and the persistent constraint displayed in Figure 6 suggests that AI can widen access to evidence-based reasoning while still leaving the most cognitively demanding interpretive work insufficiently supported. Simple tools such as evidence–claim tables, traceable AI prompts, and short feedback cycles may therefore be necessary to make reasoning more visible and help students connect source information to defensible conclusions more consistently (Kasneci et al., 2023; Hattie & Timperley, 2007).

5. CONCLUSION

This comparative study shows that AI-supported scaffolding can improve students' evidence-based critical thinking in a STEAM microcontroller media-design project. Students in the AI-supported class performed better overall than those in the non-AI class, with the strongest gains appearing in journal abstract interpretation

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

and hypothesis and reasoning, suggesting that AI was most helpful when students needed to process evidence, connect ideas, and justify design decisions. Positive effects were found across different performance levels, with the greatest benefit among lower-to-mid performers, indicating that AI can support a broad range of learners while offering particular value for students who need more guidance during complex reasoning tasks. Journal abstract interpretation nevertheless remained the main bottleneck in both groups, meaning that AI reduced the severity of this difficulty but did not remove it entirely. Practical implications follow from this pattern: teachers should use AI as a structured support tool for tasks such as evidence searching, abstract reading, claim–evidence linking, and explanation revision, while requiring students to verify AI outputs and justify conclusions in their own words. These findings should be interpreted as strong comparative evidence rather than definitive causal proof because the study used a non-randomized intact-class design. Overall, AI appears most effective when it supports students' reasoning processes rather than replacing them.

6. REFERENCES

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education, 3*, Article 60. https://doi.org/10.1007/s44217-024-00138-2

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26–55. https://doi.org/10.1002/sce.20286

Blikstein, P. (2013). Digital fabrication and "making" in education: The democratization of invention. In J. Walter-Herrmann & C. Büching (Eds.), *FabLabs: Of machines, makers and inventors* (pp. 1–22). Transcript Verlag. https://share.google/76G60acxkrzaFKJsi

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD. https://share.google/1BgX7tSXmRR8pckeS

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of Educational Research, 80*(3), 336–371. https://doi.org/10.3102/0034654310376953

Chernozhukov, V., Fernández-Val, I., & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica, 81*(6), 2205–2268. https://doi.org/10.3982/ECTA10582

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dignath, C., Büttner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review, 3*(2), 101–129. https://doi.org/10.1016/j.edurev.2008.02.003

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice, 32*(3), 179–186. https://doi.org/10.1080/00405849309543594

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report)*. American Philosophical Association. https://share.google/V6BILm5KABRlwY3Db

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica, 75*(1), 259–286. https://doi.org/10.1111/j.1468-0262.2007.00738.x

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press. https://doi.org/10.4324/9781315885278

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. https://share.google/Em6IhOHThMk1kV4Vm

Honey, M., Pearson, G., & Schweingruber, H. (Eds.). (2014). *STEM integration in K–12 education: Status, prospects, and an agenda for research*. National Academies Press. https://share.google/Xe8Y8R8v2usXbmX3X

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1913643

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

McNeill, K. L., & Krajcik, J. (2008). Inquiry and scientific explanations: Helping students use evidence and reasoning. *Science Education, 92*(2), 223–254. https://doi.org/10.1002/sce.20205

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://share.google/iCVajQWmDPdVIA9NW

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994–1020. https://doi.org/10.1002/tea.20035

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books. https://share.google/r62JMhoM89licb8b2

Paul, R., & Elder, L. (2006). *The miniature guide to critical thinking: Concepts & tools* (4th ed.). Foundation for Critical Thinking. https://share.google/Sijbdb8redRwOfBd0

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301_2

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698

Thomas, J. W. (2000). *A review of research on project-based learning*. Autodesk Foundation. https://share.google/n9PuFKu4bh6ZLHGnF

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://share.google/057C8b8YaNYkI0wre

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. https://share.google/UBqDTqbkWcTD3H3f9

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://share.google/IlYf9CvWuhmioBv5V

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v10i1.115698