International Journal of Pedagogy and Teacher Education

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

Effectiveness of STEM-Based Differentiated Learning in Enhancing Elementary Students Critical Thinking Skills

Dewi Kristyowening*, Idam Ragil Widianto Atmojo, Matsuri Matsuri
Master's Program in Elementary School Teacher Education, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Surakarta, Indonesia

ARTICLE INFO

Article History
Received :,March 11, 2025
1st Revision : April 17, 2025
Accepted : July 02, 2025
Available Online: October 30, 2025

Keywords:
STEM;
differentiated instruction;
critical thinking;
elementary education

Corresponding Author
Email address:
idamragil@fkip.uns.ac.id

ABSTRACT

Developing critical thinking is essential in contemporary schooling to prepare students to analyze, evaluate, and solve complex problems. This study examined the effectiveness of STEM-based differentiated instruction in enhancing the critical thinking of elementary students. A quasi-experimental design assigned intact classes to an experimental group receiving STEM-differentiated lessons and a control group receiving traditional cooperative instruction. Critical thinking was assessed through essay-type pretests and posttests, which were aligned with analytic rubrics. Independent *t*-tests revealed significantly greater gains for the experimental group compared to the control ($t = 11.76$, $p = 1.99 \times 10^{-17}$). N-Gain analysis revealed a moderate improvement in the experimental group (0.47) compared to a low improvement in the control group (0.19). These results suggest that integrating differentiation strategies—readiness-based tasks, choice of process and product, and flexible grouping—within STEM contexts strengthens conceptual understanding, increases cognitive engagement, and fosters inquiry-driven exploration tailored to individual needs. The findings support the pedagogical value of STEM-based differentiated instruction for cultivating higher-order thinking in the elementary grades. Practical implications include sustained teacher professional development, curriculum resources that enable tiered tasks and authentic problems, and supportive scheduling to facilitate iterative design and reflection. Future research should examine longer-term retention, classroom implementation fidelity, and differential effects across student subgroups.

How to cite: Kristyowening, D., Atmojo, I. R. W., & Matsuri, M. (2025). Effectiveness of STEM-based differentiated learning in enhancing elementary students' critical thinking skills. *International Journal of Pedagogy and Teacher Education, 9*(2), 339–349. https://doi.org/10.20961/ijpte.v9i2.100389

1. INTRODUCTION

The Fourth Industrial Revolution has accelerated technological and social change, pressing education systems to cultivate core twenty-first-century competencies—critical thinking, creativity, collaboration, and communication (Fajari, 2020; Nababan, 2019). Among these, critical thinking warrants explicit attention in the earliest grades because it underpins students' capacity to analyze, evaluate, and solve complex problems in everyday contexts (Mutakinati, 2018; Puspita & Aloysius, 2019). Evidence from international benchmarks suggests that Indonesian students' critical thinking remains relatively weak compared to their regional peers (Changwong, 2018; Nababan, 2019). The 2022 Programme for International Student Assessment (PISA) reported that Indonesia's average mathematics score was 379, substantially below the OECD average of 489 (OECD, 2023). This score represents a 15-point decline from 2018, approximating the loss of nearly three-quarters of a year of learning (OECD, 2023). Taken together, these indicators highlight persistent difficulties in applying critical thinking and intensify the urgency of systematic instructional responses from the primary years onward.

Several instructional factors have been identified as contributing to these outcomes within the Indonesian elementary education context. Classroom practices frequently prioritize memorization and coverage over deep conceptual understanding, thereby constraining opportunities for analytical engagement and transfer (Fajari, 2020; Fuad, 2017). Limited teacher expertise in inquiry-based and problem-solving pedagogies further sustains teacher-centered routines, which provide few opportunities for students to reason, argue, and reflect (Saragih & Zuhri, 2019; Setiawati & Corebima, 2017). Assessment practices often privilege recall, which signals to students that surface learning is sufficient and reduces incentives to engage in higher-order thinking. These patterns collectively impede the development of metacognitive regulation and the dispositions associated with critical thinking, such as open-mindedness and intellectual perseverance. Consequently, a pedagogical shift

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

toward active, cognitively demanding learning experiences is required to close the observed gaps in critical thinking.

STEM-based differentiated instruction has emerged as a promising approach to address these challenges by tailoring content, process, and product to students' readiness and learning profiles within integrated science, technology, engineering, and mathematics contexts (Fajari, 2020; Saragih & Zuhri, 2019; Gheyssens et al., 2020). In this model, readiness-based grouping, tiered tasks, and student choice are leveraged to align cognitive demand with prior knowledge while maintaining access to shared conceptual goals. Project-based and authentic problem-solving activities situate learning in meaningful contexts that require analysis, inference, explanation, and evaluation—core indicators of critical thinking (Mutakinati, 2018). The approach is grounded in constructivist learning theory, emphasizing knowledge construction through inquiry, collaboration, and guided discovery (Vygotsky, 1978). Scaffolding enables learners to operate within their zones of proximal development, gradually internalizing disciplinary reasoning practices as supports are faded. When implemented with fidelity, differentiated STEM learning has the potential to increase cognitive engagement, strengthen conceptual coherence, and sustain motivation over extended investigations (Fajari, 2020; Mutakinati, 2018).

Despite these theoretical and empirical promises, research that explicitly combines differentiation and STEM at the primary level remains limited. Many studies examine general STEM implementations or focus on secondary education, leaving open questions about the appropriateness of these approaches for younger learners and their alignment with cognitive development (Fajari, 2020; Saragih & Zuhri, 2019). Moreover, relatively few investigations situate interventions within the Indonesian policy context, including the *Kurikulum Merdeka*, which advocates flexible, student-centered learning. Methodologically, prior work has often lacked comprehensive assessment tools capable of capturing multi-dimensional facets of critical thinking in elementary students. There is therefore a need for studies that both adapt instructional models to learner characteristics and employ valid, reliable measures linked to established frameworks. Addressing these gaps would provide clearer guidance for teacher professional development, curriculum design, and resource allocation in schools (Nababan, 2019; Widjajarto et al., 2024; Sarwanto, 2021).

The present study addresses these needs by evaluating the effectiveness of STEM-based differentiated instruction in improving elementary students' critical thinking and by developing an assessment instrument aligned with Facione's framework, which targets analysis, inference, explanation, and evaluation (Facione, 2020). Using a quasi-experimental design that contrasts differentiated STEM instruction with cooperative learning, the research examines whether the former produces greater gains in critical thinking. The study simultaneously investigates how readiness- and profile-based differentiation strategies shape students' engagement in STEM-oriented inquiry activities. By integrating instructional design with measurement innovation, the study aims to generate actionable evidence for scaling adaptive STEM pedagogy in Indonesian elementary schools. The anticipated contributions include empirically grounded insights into effectiveness, practical guidance for implementation under typical resource constraints, and implications for assessment practices aligned with national reforms. Ultimately, the study aims to inform policy and practice by demonstrating how differentiated STEM instruction can support the complex cognitive skills essential for success in the twenty-first century.

2. MATERIAL AND METHOD

Research Design and Participants

This study employed a quasi-experimental nonequivalent-groups design to evaluate the effectiveness of STEM-based differentiated instruction in improving elementary students' critical thinking (Campbell & Stanley, 1963). The design was selected because full random assignment at the school level was infeasible, which is a common constraint in authentic school settings. Two intact groups were used: an experimental group that received STEM-based differentiated instruction and a control group that received conventional cooperative learning aligned with the standard curriculum. To strengthen internal validity, both groups completed identical pretests and posttests, allowing for analysis of within-group change and between-group contrasts. Potential threats related to selection, maturation, and history were mitigated through comparable school selection, synchronized schedules, and equal instructional time. The design notation is summarized in Table 1 to ensure transparency and replicability of the research protocol.

The study population comprised all Grade VI students across nine public elementary schools in the Ngurah Rai Cluster, Wonogiri Regency (N = 123). Cluster random sampling was applied to select schools, following

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

established guidance for educational sampling in clustered populations (Sugiyono, 2017). In line with Campbell and Stanley's (1963) recommendation that samples should comprise at least 20% of the population, four schools with similar accreditation status and enrollment size were included. The experimental group consisted of Public Elementary School 1 Gondang (SD Negeri 1 Gondang; 22 students) and Public Elementary School 1 Tegalrejo (SD Negeri 1 Tegalrejo; 12 students). The control group consisted of Public Elementary School 2 Biting (SD Negeri 2 Biting; 22 students) and Public Elementary School 1 Kepyar (SD Negeri 1 Kepyar; 12 students). Assignment at the school level minimized contamination between conditions and preserved ecological validity by maintaining intact classes

Table 1. Research Design

Group	Pretest	Treatment	Posttest
Experiment	O1	X	O2
Control	O3	Y	O4

Information:

- O1 and O3 are the pretest results for the experimental and control groups.
- X is the treatment given to the experimental group.
- Y is the treatment or condition given to the control group (usually no special treatment or standard treatment).
- O2 and O4 were posttest results for the experimental and control groups.

Instruments and Measurements

The principal outcome measure comprised six essay prompts aligned with the Grade 6 science topic "Rotation and Revolution of the Earth." Item construction was guided by Facione's critical thinking framework, targeting analysis, inference, explanation, and evaluation at an elementary-appropriate level (Facione, 2020). The prompts were adapted to elicit skills such as identifying causal relationships, comparing natural phenomena, justifying judgments with evidence, and applying concepts to real-life situations. An analytic rubric mapped to the four indicators was used to ensure consistent scoring across occasions and raters. Prior to implementation, the instrument underwent expert review for content relevance and cognitive appropriateness, followed by a pilot to refine wording, difficulty, and scoring guidance. Complementing the essay test, an 18-item Likert-type questionnaire measured students' perceptions of their critical thinking opportunities during instruction, including perceived challenge, clarity, and engagement. Item validity was examined using Pearson product–moment correlations, and all items met the inclusion threshold. Reliability analysis yielded Cronbach's alpha = 0.62, indicating acceptable internal consistency for exploratory educational research with this age group.

To support criterion-referenced interpretation, an interpretation scale (ri) was employed to classify performance levels across relevant indices. Five ordered categories—Very Low, Low, Moderate, High, and Very High—were specified a priori to enhance transparency and minimize post hoc bias. The scale was used descriptively to summarize distributions and to complement inferential statistics with actionable proficiency levels. Such categorical reporting facilitates communication to practitioners and policymakers who require thresholds rather than only mean differences and *p*-values. The categories and value ranges are presented in Table 2 and were applied consistently across analyses. Notably, this table represents an interpretation scale rather than demographic information, thereby correcting the earlier mislabeling.

Table 2. Interpretation Scale (ri)

Criterion	Value Range (ri)
Very High	$0.80 < ri \leq 1.00$
High	$0.60 < ri \leq 0.80$
Moderate	$0.40 < ri \leq 0.60$
Low	$0.20 < ri \leq 0.40$
Very Low	$0.00 < ri \leq 0.20$

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

The study followed a structured sequence comprising a pretest, treatment, and posttest phase to estimate the change attributable to the intervention. The experimental group received a six-week STEM-based differentiated learning program delivered in two 60-minute sessions per week, ensuring adequate exposure while aligning with school schedules. Instruction in the experimental condition was problem-based and inquiry-oriented, with systematic differentiation of content, process, and product according to students' readiness levels and learning profiles. The control group participated in conventional instruction based on cooperative learning aligned with the standard curriculum, thereby providing a credible business-as-usual comparison. All sessions were delivered by classroom teachers who had been trained in advance and supervised by the research team to promote fidelity of implementation. Pretests were administered immediately before the intervention, and posttests were administered immediately after the final session, using standardized protocols across schools to ensure comparability and scoring integrity.

Data collection

Data were collected through three staged techniques designed to yield complementary quantitative and qualitative evidence for evaluating the intervention. The sequence comprised (a) pretest and posttest administrations, (b) systematic classroom observation, and (c) an end-of-unit student questionnaire. The staged design enabled the study to establish baseline performance, document instructional processes, and capture learners' perceptions in a coherent manner. Standardized protocols, scripts, and timing were applied across schools to promote procedural consistency and reduce instrumentation bias. Multiple sources were purposefully combined to enable triangulation, thereby strengthening the credibility of inferences drawn from the findings. All procedures adhered to school policies and were implemented by trained personnel under the supervision of the research team.

The first stage involved administering pretests and posttests to assess changes in students' critical thinking resulting from the intervention. The pretest was administered immediately before instruction began, and the posttest was administered after the final session, bracketing the six-week period. Tests were delivered by classroom teachers who had been briefed in advance using a common script to ensure uniform instructions and time limits across sites. Test content aligned with the study's indicators of analysis, inference, explanation, and evaluation, and corresponded to the Grade 6 topic "Rotation and Revolution of the Earth." Administrative procedures emphasized standardization and test security, including the use of consistent materials and testing conditions for all students. Resulting scores formed the basis for between-group comparisons and for estimating learning gains, which were subsequently analyzed using the statistical procedures described in the Data Analysis section.

The second stage involved structured observations of classroom instruction to document the enactment of STEM-based differentiated instruction and the comparison condition. Researchers served as nonparticipant observers to minimize disruption and preserve typical classroom dynamics during data collection. A validated observation protocol guided evidence gathering, focusing on (1) clarity of teacher instructions, (2) levels of active student participation, and (3) the smooth implementation of lesson phases. Observers also recorded instances of inquiry, problem-solving discourse, and differentiation moves, including tiered tasks and flexible grouping where present. Observation visits were scheduled across schools and weeks to provide a balanced and representative sample of lessons in both conditions. Field notes and checklist ratings were later synthesized to contextualize test results, assess implementation fidelity, and support triangulation with quantitative outcomes.

The final stage consisted of administering a Likert-type questionnaire to students upon completion of the instructional period. The instrument elicited perceptions of learning comfort, clarity of materials and instructions, engagement in activities, and the perceived usefulness of STEM-oriented tasks for cultivating critical thinking. Administration occurred during class time under teacher supervision and followed a standardized script to maintain procedural reliability across schools. Students completed the questionnaire individually, after being assured of confidentiality, to encourage candid responses about their learning experiences. Items were aligned with the study's indicators of analysis, inference, explanation, and evaluation to enable coherent interpretation alongside test scores. The resulting data were used descriptively to characterize students' experiences and to enrich the interpretation of the pretest–posttest outcomes through convergent evidence.

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

Data analysis

Data analysis proceeded in three consecutive stages using IBM SPSS Statistics, Version 25, to ensure rigor and reproducibility. The first stage comprised assumption testing to determine the appropriateness of parametric inference for the study variables. Specifically, distributional normality was examined using the Shapiro–Wilk test for each group at both the pretest and posttest, and homogeneity of variances was assessed with Levene's test. These tests were applied to the composite critical-thinking scores to verify that both experimental and control groups met the assumptions required for *t*-based comparisons. Assumption checks were conducted prior to any inferential testing and were documented for transparency. When the assumptions were satisfied, parametric procedures were retained; when assumptions were marginal, analyses were corroborated with robust summaries to confirm stability of the conclusions. This multi-step approach strengthened internal validity by aligning analytic choices with observed data characteristics.

The second stage focused on evaluating the quality of the study instruments to support trustworthy inferences. Item validity for the questionnaire was assessed using Pearson product–moment correlations, and items meeting the pre-established criterion were retained for further analysis. Internal consistency reliability was estimated using Cronbach's alpha, yielding a coefficient of 0.62, which is acceptable for exploratory educational research with elementary-aged participants. The essay-based measure was scored using an analytic rubric aligned to the study's indicators of analysis, inference, explanation, and evaluation, thereby enhancing content validity. Expert review and piloting further supported the alignment between items and targeted constructs prior to full implementation. Together, these procedures provided evidence that the measurement tools were suitable for capturing changes in students' critical thinking within the study context. Instrument diagnostics and decision rules were archived to ensure replicability and auditability of the analytic workflow.

Table 3. Critical Thinking Criteria

Score	Criterion
81.25 < x ≤ 100	Very High
71.50 < x ≤ 81.25	High
62.50 < x ≤ 71.50	Moderate
43.75 < x ≤ 62.50	Low
0.00 < x ≤ 43.75	Very Low

The third stage estimated the effect of the instructional intervention on students' critical thinking outcomes. Primary comparisons employed independent-samples *t*-tests on posttest scores to evaluate between-group differences while reporting pretest statistics to contextualize baseline equivalence. Statistical significance was evaluated using two-tailed tests at $\alpha = 0.05$, and decisions were based on *p*-values rather than "*t* table" lookups to align with contemporary reporting standards. In addition to null-hypothesis tests, effect sizes (e.g., Cohen's *d* using pooled standard deviations) were planned to quantify the magnitude of observed differences and to facilitate interpretation of practical significance. Normalized gain (N-gain) was also computed to summarize improvement relative to the maximum possible gain, complementing mean differences with an interpretable growth metric. Confidence intervals for mean differences and effect sizes were reported to convey precision and support evidence-based conclusions. Collectively, these steps provided convergent statistical evidence regarding the effectiveness of STEM-based differentiated instruction. To aid criterion-referenced interpretation, essay scores from the pretest and posttest were classified into performance levels using a rubric adapted from Karim and Normaya (2015). The rubric distinguishes five ordered categories—Very Low, Low, Moderate, High, and Very High—based on percentage score ranges. Such categorical reporting enables practitioners and policymakers to interpret outcomes beyond mean scores by identifying the proportion of students who meet or exceed meaningful thresholds. Classification was applied consistently at both time points to allow comparisons of distributional shifts attributable to the intervention. The categorical scheme complements inferential statistics by offering an accessible summary of student proficiency. Table 3 presents the thresholds used for classification, which were applied uniformly across both study groups.

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

3. RESULTS

Assumption Checks: Normality and Homogeneity

Establishing that the data satisfied parametric assumptions was a prerequisite for the inferential analyses reported below. As summarized in Table 4, the distributions of pretest and posttest scores for both the control and experimental groups did not deviate significantly from normality (all *p*-values > 0.05). These findings indicate that the observed score patterns are compatible with the normal model typically assumed by *t*-based procedures and related parametric tests. The consistency of conclusions across measurement occasions strengthens confidence that later comparisons are not artifacts of skewed or heavy-tailed distributions. Importantly, the use of a common testing protocol across schools further reduced potential procedural sources of non-normality. Together, the evidence in Table 4 supports the analytic decision to proceed with parametric testing for group contrasts. This decision set the stage for subsequent checks of variance equality and the estimation of treatment effects.

Variance homogeneity was examined using Levene's test to ensure comparability of dispersion across groups at each time point. As reported in Table 5, pretest and posttest comparisons both yielded non-significant results (all *p*-values > 0.05), indicating that the equal-variances assumption was satisfied. Meeting this assumption reduces the risk of biased standard errors and maintains the nominal Type I error rate for independent-samples *t*-tests. The joint satisfaction of normality and homogeneity assumptions supports the validity of mean-based inferential comparisons between conditions. In addition, these diagnostics indicate that any subsequent differences are unlikely to be driven by heteroscedasticity or irregular distributional shapes. Accordingly, the analysis proceeded to instrument checks, descriptive summaries, and hypothesis testing using the planned parametric approach. Throughout, Tables 4–5 are cited as the empirical basis for these decisions.

Table 4. Population Normality Test Results

Class	Statistics	p-value	Conclusion
Control - Pretest	0.9808	0.8219	Usual
Control - Posttest	0.9477	0.1237	Usual
Experiment - Pretest	0.9618	0.3075	Usual
Experiment - Posttest	0.9867	0.9547	Usual

Table 5. Population Homogeneity Test Results

Test	Statistics	p-value	Conclusion
Pretest Homogeneity	0.0044	0.9475	Homogeneous
Posttest Homogeneity	0.9869	0.3244	Homogeneous

Instrument Quality: Validity and Reliability

The primary assessment consisted of six essay items aligned with the Grade 6 topic "Rotation and Revolution of the Earth" and operationalized under Facione's indicators of analysis, inference, explanation, and evaluation. Item screening demonstrated that all prompts met the validity criterion and were retained for operational use, as shown in Table 6. This outcome indicates adequate item–construct alignment and supports the interpretability of total and domain scores in subsequent analyses. Internal consistency for the questionnaire was estimated with Cronbach's alpha, yielding $\alpha = 0.62$, which is acceptable for exploratory educational research with elementary-aged learners. The combination of expert review, piloting, and statistical screening provides convergent evidence for instrument adequacy in this context. Moreover, the instrument's focus on multiple indicators of critical thinking increases content coverage and reduces construct under-representation. Collectively, the evidence in Table 6 justifies the use of the instrument to detect change attributable to the intervention. To promote scoring consistency, responses to the essay items were evaluated with an analytic rubric explicitly mapped to the four critical-thinking indicators. Rubric descriptors specified criteria for evidence, reasoning, and clarity, thereby standardizing expectations across raters and occasions. This design choice reduces

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

measurement error associated with subjective scoring and supports reliable aggregation to composite indices. The rubric also facilitates criterion-referenced interpretation, enabling subsequent classification of scores into performance levels reported elsewhere in the manuscript. In parallel, questionnaire items elicited perceptions of challenge, clarity, and engagement, aligning self-report data with the same conceptual indicators used in performance scoring. Such alignment aids triangulation between perceived opportunities for thinking and demonstrated outcomes on the essays. Taken together, these procedures enhanced the validity and reliability of inferences drawn from the assessment battery, as documented in Table 6.

Table 6. Question Validation Results

Criterion	Question Number	Sum
Valid	1, 2, 3, 4, 5, 6	6

Descriptive Statistics and Learning Gains

Group means for pretest and posttest administrations are presented in Table 7 to provide an initial picture of performance patterns. The experimental group's mean increased from 62.00 at pretest to 80.00 at posttest, reflecting an 18-point gain over the intervention period. By contrast, the control group's mean rose from 60.00 to 70.00, indicating a 10-point gain under business-as-usual instruction. The larger absolute increase for the experimental group suggests that STEM-based differentiated instruction supported greater improvement in critical-thinking performance. Because both groups experienced identical testing windows and instructional time, these descriptive differences are unlikely to be attributable solely to exposure. The direction and magnitude of changes are consistent with the study's theory of action, which posits that inquiry-rich, differentiated tasks should elevate analytical engagement. Overall, Table 7 provides descriptive support for a treatment advantage prior to formal hypothesis testing. To contextualize improvement relative to potential headroom, normalized gain (N-gain) was computed as *(posttest − pretest) / (ideal − pretest)*. As shown in Table 8, the experimental group achieved an average N-gain of 0.47, classified as "Moderate," whereas the control group achieved 0.19, classified as "Low." These categorical interpretations highlight not only absolute growth but also proportional progress toward the maximum attainable score. The contrast in N-gain indicates that students in the experimental condition converted a larger share of available learning opportunity into actual performance gains. For ease of comparison, Figure 1 visualizes the disparity, making the distinction between moderate and low values apparent at a glance. The pattern in Figure 1 reinforces the interpretation that the intervention resulted in significantly greater learning efficiency. Together, Tables 7–8 and Figure 1 provide convergent descriptive evidence in favor of the STEM-based differentiated approach.

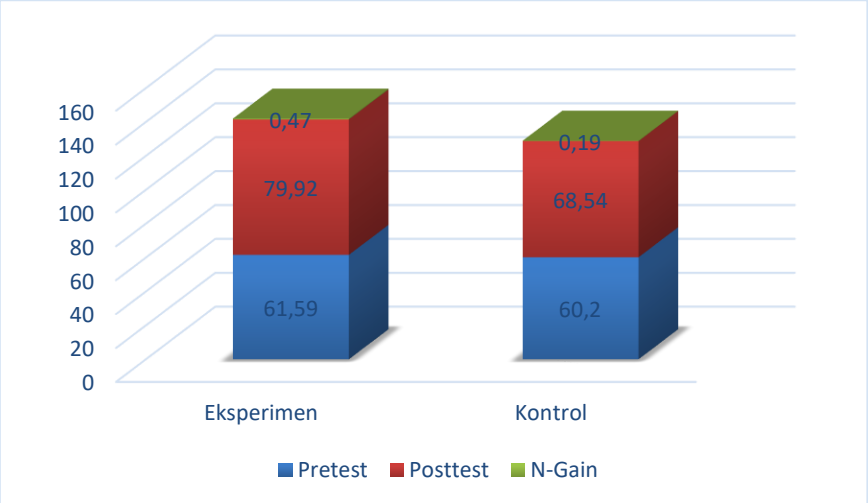

Figure 1. Graph of N-Gain Critical Thinking

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

Table 7. Mean Pretest and Posttest Scores by Group

Score	Experimental Classes	Control Classes
Pretest	62.00	60.00
Posttest	80.00	70.00

Table 8. N-Gain Test Results

Class	N-Gain	Category
Experiment	0.47	Moderate
Control	0.19	Low

Hypothesis Testing and Practical Significance

Between-group differences at posttest were evaluated using an independent-samples *t*-test to test the primary efficacy hypothesis. Results in Table 9 show a statistically significant advantage for the experimental group, $t(62) = 11.76$, with $p = 1.99 \times 10^{-17}$, exceeding the conventional $\alpha = 0.05$ threshold by a wide margin. The reported *t*-value also exceeds the corresponding critical value (1.66), corroborating the rejection of the null hypothesis under standard decision rules. These findings indicate that the observed posttest difference is highly unlikely to have arisen by chance given the sample sizes and variances. The robustness of the result is consistent with the descriptive patterns established earlier for means and gains. In short, Table 9 provides strong inferential evidence that STEM-based differentiated instruction improved critical-thinking performance relative to the comparison condition. This conclusion directly addresses Research Question 1 and provides a statistical foundation for discussing educational implications.

Complementary descriptive statistics in Table 10 further illuminate the nature of the treatment effect. The experimental group obtained a mean of 79.92 (SD = 3.80), whereas the control group obtained a mean of 68.54 (SD = 3.94), indicating both a higher central tendency and comparable dispersion under the intervention. The standardized mean difference, Cohen's $d = 2.91$, signifies a very large effect size, suggesting that the magnitude of benefit is not only statistically significant but also educationally substantial. Similar standard deviations across groups imply relatively uniform gains among students exposed to the intervention rather than improvements confined to a small subset. This distributional pattern aligns with the moderate N-gain classification for the experimental group and the low N-gain classification for the control group (see Tables 7–8 and Figure 1). The coherence of descriptive, gain-based, and inferential results strengthens the conclusion that the intervention produced meaningful and practically important improvements. Collectively, Tables 9–10 document the statistical and practical significance of the treatment effect on critical-thinking outcomes.

Table 9. Independent Samples t-Test Results: Posttest Scores

Comparison	t-value	df	t-table (α = 0.05)	p-value	Conclusion
Experimental vs. Control Posttest	11.76	62	1.66	1.99×10^{-17}	Experimental group significantly outperformed control

Table 10. Descriptive Statistics for Posttest Scores

Group	N	Mean	Std. Deviation
Experimental	32	79.92	3.80
Control	32	68.54	3.94

4. DISCUSSION

Effectiveness of STEM-Based Differentiated Instruction versus Cooperative Learning

The findings provide clear evidence that STEM-based differentiated instruction yields superior gains in critical thinking relative to conventional cooperative learning. Descriptively, the experimental group outperformed the control group on the posttest, with larger mean growth from pretest to posttest (see Table 7)

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

and proportionally greater improvement as captured by N-Gain (see Table 8). This pattern is visualized in Figure 1, where the experimental group's moderate N-Gain (0.47) contrasts with the control group's low N-Gain (0.19), indicating more efficient learning relative to the available headroom. Inferentially, the independent-samples *t*-test confirmed a large and statistically significant difference at posttest, $t(62) = 11.76$, $p < 0.001$, favoring the experimental condition (see Table 9). Complementary descriptive statistics underscore the magnitude and stability of the effect, with higher means and comparable dispersion in the experimental group (see Table 10). The resulting Cohen's *d* of 2.91 indicates a very large practical impact, suggesting that the treatment effect is not only statistically robust but also educationally consequential.

These conclusions are strengthened by the study's diagnostic checks and the quality of its measurements. Normality and homogeneity assumptions were satisfied for all relevant comparisons, supporting the appropriateness of parametric analyses and protecting against inflated Type I error (see Table 4 and Table 5). Moreover, instrument screening and reliability evidence indicated that the assessment tools were suitable for detecting change in the targeted constructs of analysis, inference, explanation, and evaluation (see Table 6). The convergence of descriptive, inferential, and gain-based indicators across Tables 7–10 and Figure 1 provides a coherent answer to RQ1: STEM-based differentiated instruction significantly improves elementary students' critical-thinking performance compared to cooperative learning. These results align with prior reports of sizable effects for STEM-oriented pedagogies (Haetami, 2023; Noufal, 2022), thereby situating the present findings within a growing evidence base that links inquiry-rich, problem-oriented instruction to higher-order learning outcomes. Taken together, the pattern supports both the statistical and practical superiority of the treatment condition for cultivating critical thinking in Grade 6 science.

Engagement Mechanisms of Readiness- and Profile-Based Differentiation in STEM Inquiry

The observed advantages are consistent with theoretically grounded mechanisms through which differentiation can heighten engagement in STEM inquiry. By tailoring content, process, and product to students' readiness and learning profiles, the intervention likely calibrated cognitive demand, provided accessible entry points, and sustained productive struggle—conditions known to promote participation and persistence (Tomlinson, 2021; Sen et al., 2021; Shanta & Wells, 2022). Within this structure, project- and problem-based tasks afforded authentic contexts for analysis, evaluation, and transfer, while flexible grouping and scaffolded supports aligned with Vygotskian notions of assisted performance and the zone of proximal development (Vygotsky, 1978). Although engagement was not the primary outcome, the moderate N-Gain observed in the experimental group (see Table 8 and Figure 1) is consistent with heightened cognitive engagement that translates available learning opportunities into realized performance. This interpretation also accords with literature showing that collaborative, inquiry-rich STEM environments can amplify critical-thinking growth when coupled with differentiated task structures (Mater et al., 2020). In short, the readiness- and profile-based features appear to have functioned as levers that increased time-on-task quality and depth of processing during STEM inquiry.

The implications and boundaries of these mechanisms merit careful consideration for practice and future study. For implementation, the results suggest prioritizing teacher professional development in tiering, flexible grouping, and rubric-guided scaffolding so that differentiation reliably targets analysis, inference, explanation, and evaluation—the very indicators assessed in this study (see Table 6). Schools should also ensure access to multimodal resources and scheduling structures that support iterative design cycles, as these conditions likely contributed to the observed gains (see Tables 7–10 and Figure 1). At the same time, limitations—such as the study's single-district sample, potential variability in implementation fidelity, and reliance on essay-based and self-report measures—constrain broad generalization. Future research should examine diverse contexts, track longer-term retention, and incorporate performance-based and observational engagement metrics to directly test the proposed mechanisms. Investigating moderators such as teacher expertise, classroom climate, and family involvement would clarify when and for whom differentiation most powerfully enhances STEM inquiry. By addressing these directions, subsequent studies can more precisely connect readiness- and profile-based design decisions to measurable engagement patterns and sustained critical-thinking development.

5. CONCLUSION

This study demonstrates that STEM-based differentiated instruction significantly enhances elementary students' critical thinking relative to conventional approaches. Students who received instruction tailored to their

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

readiness and learning profiles within an integrated STEM framework showed greater gains in analysis, evaluation, and problem-solving than their peers in the comparison group. These results align with constructivist principles, indicating that active, student-centered environments foster deeper cognitive engagement and more durable conceptual understanding. By calibrating task complexity, process, and product to learners' profiles, differentiation created accessible entry points while sustaining productive challenge across the unit. The approach also appears to support transfer, as students applied concepts to novel, real-world problems that required coordinated reasoning across science, technology, engineering, and mathematics. Collectively, the evidence affirms the pedagogical value of combining STEM inquiry with systematic differentiation in the elementary grades. The findings carry several practical implications for curriculum and instruction. Schools should invest in sustained professional development that equips teachers to implement tiered tasks, flexible grouping, and rubric-guided scaffolding aligned with critical-thinking indicators. Resource provision—such as manipulatives, technological tools, and time for iterative design cycles—will further enable high-fidelity enactment of differentiated STEM lessons. Assessment systems should be aligned to the targeted constructs, pairing analytic rubrics with opportunities for inquiry-based performance to capture growth beyond recall. Integrating inquiry-driven problem-solving into routine STEM teaching can yield more meaningful learning experiences and promote cross-disciplinary application of critical-thinking skills. Future research should examine scalability across diverse contexts, monitor longer-term retention, and investigate moderators such as teacher expertise and classroom climate. Taken together, these steps can help realize the full potential of STEM-based differentiated instruction to cultivate higher-order thinking from the earliest grades.

6. REFERENCES

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Rand McNally.

Changwong, K., Sukkamart, A., & Sisan, B. (2018). Critical thinking skill development: Analysis of a new learning management model for Thai high schools. *Journal of International Studies, 11*(2), 37–48. https://doi.org/10.14254/2071-8330.2018/11-2/3

Facione, P. A. (2020). *Critical thinking: What it is and why it counts (2020 update).* Insight Assessment. https://insightassessment.com/wp-content/uploads/2023/12/Critical-Thinking-What-It-Is-and-Why-It-Counts.pdf

Fajari, L. E. W., Sarwanto, & Chumdari. (2020). The effect of problem-based learning multimedia and picture media on students' critical-thinking skills viewed from learning motivation and learning styles in elementary school. *Elementary Education Online, 19*(3), 1797–1811. https://doi.org/10.17051/ilkonline.2020.735165

Fuad, N. M., Zubaidah, S., Mahanal, S., & Suarsini, E. (2017). Improving junior high schools' critical thinking skills based on test three different models of learning. *International Journal of Instruction, 10*(4), 101–116. https://doi.org/10.12973/iji.2017.1017a

Gheyssens, E., Consuegra, E., Engels, N., & Struyven, K. (2020). Good things come to those who wait: The importance of professional development for the implementation of differentiated instruction. *Frontiers in Education, 5*, 96. https://doi.org/10.3389/feduc.2020.00096

Haetami, H. (2023). Effect of STEM-based differentiated learning to improve students' critical thinking skills: A meta-analysis study. *Jurnal Penelitian Pendidikan IPA, 9*(9), 655–663. https://doi.org/10.29303/jppipa.v9i9.5084

Karim, N., & Normaya. (2015). Kemampuan berpikir kritis siswa dalam pembelajaran matematika dengan menggunakan model JUCAMA di sekolah menengah pertama. *EDU-MAT: Jurnal Pendidikan Matematika, 3*(1), 92–104. https://doi.org/10.20527/edumat.v3i1.634

Mater, N., Hussein, M. H., Salha, S., Draidi, F., Shaqour, A., Qatanani, N., & Affouneh, S. (2020). The effect of the integration of STEM on critical thinking and the technology acceptance model. *Educational Studies, 48*, 642–658. https://doi.org/10.1080/03055698.2020.1793736

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389

Mutakinati, L., Anwari, I., & Yoshisuke, K. (2018). Analysis of students' critical thinking skill of middle school through STEM education project-based learning. *Jurnal Pendidikan IPA Indonesia, 7*(1), 54–65. https://doi.org/10.15294/jpii.v7i1.10495

Nababan, T. S. (2019). Development analysis of global competitiveness index of ASEAN-7 countries and its relationship to gross domestic product. *International Journal of Business and Economics (IJBE), 3*(1), 1–8. https://doi.org/10.33019/ijbe.v3i1.108

Noufal, P. (2022). Effectiveness of STEM approach on enhancing critical thinking skill of secondary school students. *International Journal of Humanities, Social Sciences and Education (IJHSSE), 9*(5), 79–87. https://arcjournals.org/pdfs/ijhsse/v9-i5/8.pdf

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education.* OECD Publishing. https://doi.org/10.1787/53f23881-en

Puspita, A. S., & Aloysius, S. (2019). Developing students' critical thinking skills through implementation of problem-based learning approach. *Journal of Physics: Conference Series, 1241*, 012020. https://doi.org/10.1088/1742-6596/1241/1/012020

Saragih, S., & Zuhri, D. (2019). Teacher behavior in students' critical thinking ability development. *Journal of Physics: Conference Series, 1320*, 012006. https://doi.org/10.1088/1742-6596/1320/1/012006

Sarwanto, S. (2021). Analysis of the implementation of the independent learning curriculum (Merdeka Belajar) in science learning in elementary schools. *Indonesian Journal of Instruction, 1*(1), 37–46. https://doi.org/10.33503/iji.v1i1.37

Sen, C., Ay, Z. S., & Kiray, S. A. (2021). Computational thinking skills of gifted and talented students in integrated STEM activities based on the engineering design process: The case of robotics and 3D robot modeling. *Thinking Skills and Creativity, 42*, 100931. https://doi.org/10.1016/j.tsc.2021.100931

Setiawati, H., & Corebima, A. D. (2017). Empowering critical thinking skills of students having different academic ability in biology learning of senior high school through PQ4R–TPS strategy. *The International Journal of Social Sciences and Humanities Invention, 4*(5), 3521–3526. https://doi.org/10.18535/ijsshi/v4i5.09

Shanta, S., & Wells, J. (2022). T/E design-based learning: Assessing student critical thinking and problem-solving abilities. *International Journal of Technology and Design Education, 32*, 267–285. https://doi.org/10.1007/s10798-020-09608-8

Sugiyono. (2017). *Metode penelitian kuantitatif, kualitatif, dan R&D* (Cet. 26). Alfabeta.

Tomlinson, C. A. (2021). *So each may soar: The principles and practices of learner-centered classrooms.* ASCD. https://www.ascd.org/books/so-each-may-soar

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

Widjajarto, A., Lubis, M., & Lubis, A. R. (2024). Service level agreement (SLAs) model for disaster recovery center (DRC) based on computational resource model of virtual machine. *Procedia Computer Science, 234*, 1476–1483. https://doi.org/10.1016/j.procs.2024.03.148

p-ISSN: 2597-7792 / e-ISSN: 2549-8525
DOI: https://doi.org/10.20961/ijpte.v9i2.100389