



Implementation of the ARKA-STEM Model and Students' Mathematical Critical Thinking: An Exploratory Two-Class Quasi-Experimental Study

Darus Salam, *Dyah Worowirastri Ekowati, Agus Tinus

Master's Program in Pedagogy, Directorate of Postgraduate Programs, Universitas Muhammadiyah Malang, Malang, Indonesia

*Corresponding author e-mail: wirowirastri@umm.ac.id

Received: September 2026; Revised: October 2026; Published: October 2026

Abstract

This exploratory quasi-experimental study examined mathematical critical thinking in two intact Grade XII classes at SMAN 3 Sampit, Central Kalimantan: one taught with the ARKA (Activity, Reflection, Conceptualization, Application) model integrated with STEM ($n = 33$) and one taught with conventional direct instruction ($n = 33$), on the topic of Circle Equations and Their Elements. Because one class represented each condition, treatment is confounded with class membership; all inferential statistics are therefore reported as descriptive contrasts between these two classes and not as estimates of a population treatment effect. Critical thinking was measured with a six-item constructed-response test based on Facione's framework (one item per indicator; 0–24 scale), scored by the researcher without blinding. In the ARKA-STEM class the mean score rose from 11.06 to 17.36 (gain = 6.30, 95% CI [5.62, 6.98]; mean N-Gain = 0.510), and in the comparison class from 9.82 to 11.27 (gain = 1.45; mean N-Gain = 0.112). Adjusted for pretest, the ARKA-STEM class scored 4.96 points higher (95% CI [4.08, 5.84]); because regression slopes differed between classes ($p = .036$), the conditional difference was larger at lower pretest scores. The standardized posttest difference was $d = 1.59$ (95% CI [1.04, 2.14]). Posttest reliability in the ARKA-STEM class was low ($\alpha = .531$), so indicator-level results are exploratory. Observed implementation fidelity rose from 71.02% to 97.16% across three meetings. The ARKA-STEM class scored substantially higher than the comparison class under these conditions, but the design cannot establish that the model caused the difference; replication with several independently assigned classes, blinded scoring, and a validated instrument is required.

Keywords: ARKA model; STEM education; Critical thinking skills; Mathematics learning; Experiential learning; Facione framework

How to Cite: Salam, D., Ekowati, D. W., & Tinus, A. (2026). Implementation of the ARKA-STEM Model and Students' Mathematical Critical Thinking: An Exploratory Two-Class Quasi-Experimental Study. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, 14(4), 2376–2396. <https://doi.org/10.33394/j-ps.v14i4.22777>



<https://doi.org/10.33394/j-ps.v14i4.22777>

Copyright © 2026, Salam et al.

This is an open-access article under the [CC-BY](#) License



INTRODUCTION

The era of globalization and the Fourth Industrial Revolution has brought fundamental changes across all sectors of life, including education. Rapid advances in technology, artificial intelligence, and digitalization require human resources who not only master disciplinary content but also possess the capacity to think critically, adaptively, and innovatively. In this context, twenty-first-century education positions Higher-Order Thinking Skills (HOTS) as a central competence that every learner must acquire (González-Pérez & Ramírez-Montoya, 2022; Thornhill-Miller et al., 2023). Indonesia's national policy, articulated through Kurikulum Merdeka issued by the Ministry of Education, Culture, Research, and Technology, explicitly promotes student-centered, project-based learning oriented toward the development of holistic competencies. Among the dimensions of the Profil Pelajar Pancasila, critical reasoning occupies a pivotal place alongside creativity and collaboration (Kemdikbudristek, 2023). The

emphasis on critical reasoning reflects a broader international consensus that the ability to analyze, evaluate, and construct reasoned judgments is indispensable for active citizenship and professional success in an information-rich society.

Critical thinking is widely recognized as the core of HOTS because it enables individuals to analyze information in depth, evaluate arguments logically, identify hidden assumptions, and reach evidence-based decisions (Facione, 2020; Paul & Elder, 2020). Facione (2020) defines critical thinking as purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. The six cognitive skills that constitute this framework—interpretation, analysis, evaluation, inference, explanation, and self-regulation—provide a comprehensive operational basis for assessing critical thinking in educational research. (Paul & Elder, 2020) further emphasize the metacognitive dimension of critical thinking, arguing that skilled thinkers routinely monitor, assess, and revise their own thinking processes. Together, these conceptualizations establish critical thinking not as a single discrete skill but as an integrated constellation of cognitive and metacognitive abilities that can be cultivated through deliberate instructional design.

Mathematics is one of the most strategic school subjects for cultivating critical thinking. Far from being merely a collection of computational procedures or formula applications, mathematics demands deductive reasoning, abstraction, modeling of real-world phenomena, and systematic evaluation of solutions (Cevikbas et al., 2022; Greefrath et al., 2022; Wang & Abdullah, 2024). National Council of Teachers of Mathematics (2000) has long insisted that mathematics instruction must encourage students to communicate their mathematical thinking, justify procedures, and evaluate alternative approaches to problem solving. Within Facione's framework, mathematical critical thinking can therefore be operationalized through the same six indicators: the ability to understand and clarify mathematical data (interpretation), to identify relationships among concepts and arguments (analysis), to judge the quality and logic of solution strategies (evaluation), to draw conclusions based on available evidence (inference), to present and justify reasoning systematically (explanation), and to monitor and evaluate one's own thinking processes (self-regulation) (Yaniawati et al., 2020). Strengthening critical thinking in mathematics is thus not merely a curricular demand but a long-term investment in preparing a generation capable of meeting the complex challenges of contemporary life.

Nevertheless, empirical evidence consistently shows that Indonesian students' mathematical critical thinking remains far from satisfactory. The Programme for International Student Assessment (PISA) 2022 ranked Indonesia 68th out of 81 participating countries, with an average mathematics score of 366—well below the Organisation for Economic Co-operation and Development (OECD) average of 472 (OECD, 2023). More troubling is the finding that the majority of Indonesian students are able to solve only low-cognitive-level, routine, and procedural items, while very few succeed on items that require critical reasoning and complex problem solving. National Assessment results (Asesmen Nasional Berbasis Kompetensi / ANBK) corroborate this picture: most Indonesian students still operate at a basic level of numeracy literacy and are unable to apply multi-step mathematical reasoning to contextual problems (Kemdikbudristek, 2023). These national patterns are mirrored at the local level in East Kotawaringin Regency, Central Kalimantan, where school quality reports indicate persistently low numeracy literacy achievement (Dinas Pendidikan Kabupaten Kotawaringin, 2024). The cumulative evidence from international, national, and local assessments points to a systemic shortfall in the capacity of mathematics classrooms to develop higher-order thinking.

A preliminary study conducted by the researcher in the even semester of the 2025/2026 academic year at SMAN 3 Sampit, located in Baamang District, East Kotawaringin Regency, confirmed these difficulties in situ. Most students struggled to analyze non-routine mathematics problems, construct coherent mathematical arguments, and explore solutions

from multiple perspectives. They tended to remain passive, waiting for teacher instructions rather than constructing understanding independently. Quantitative data from the Odd Semester Summative Assessment of Grade XII students yielded a mean score of 57.8 against a mastery criterion (Kriteria Ketercapaian Tujuan Pembelajaran / KKTP) of 70, producing a deficit of 12.2 points; more than 68% of students scored below the KKTP. Error analysis of student answer sheets revealed that the most frequent difficulties were not computational but conceptual: students failed to comprehend problems comprehensively (interpretation), to identify relevant information and analyze inter-conceptual relationships (analysis), to select and justify appropriate strategies (evaluation and explanation), and to evaluate the reasonableness of obtained results (inference and self-regulation). A diagnostic critical-thinking test administered to 32 students of Grade XII Science-3 further showed that only 18.75% reached the “high” category, while 53.13% were classified as “low” and 28.12% as “moderate.” These data indicate a substantial deficit in critical thinking that requires systematic pedagogical intervention.

One of the principal factors underlying low mathematical critical thinking is the continued dominance of teacher-centered, procedural instruction. Many mathematics teachers in Indonesia still rely heavily on lectures, algorithmic demonstrations, and routine drill that demand only memorization and application of procedures, leaving little space for students to explore, discuss, pose critical questions, or construct understanding independently (Siahaan et al., 2023; Trisnani, 2024; Wang & Abdullah, 2024). Such instructional patterns systematically inhibit the development of analysis, evaluation, inference, and self-regulation—the core components of critical thinking according to (Facione, 2020). In addition, the weak connection between mathematical content and real-life contexts reduces students’ motivation to engage in deep thinking when the material appears abstract and meaningless (OECD, 2023; Schukajlow et al., 2023; Trisnani, 2024). Consequently, classroom mathematics tends to reproduce passive knowledge rather than build the active thinking competencies required in real life. Research by (Siahaan, 2023) on trends in critical-thinking studies in Indonesian mathematics education further confirms that teacher-centered approaches remain prevalent and that interventions targeting higher-order thinking are still limited in both number and geographic distribution.

Although constructivist models such as Problem-Based Learning (PBL) and Project-Based Learning (PjBL) have been widely employed to develop critical thinking, these models generally focus on the final product of problem solving or the discovery of operational concepts, often without providing structured metacognitive reflection that bridges concrete experience to deep abstract understanding. As a promising alternative, the ARKA model (Activity, Reflection, Conceptualization, and Application) developed by (Ilyasa et al., 2024) adapts the principles of David Kolb’s Experiential Learning Theory into a systematic instructional syntax. In the Activity stage, students engage in direct concrete experience; in the Reflection stage, they review and evaluate that experience; in the Conceptualization stage, they organize and formalize mathematical concepts on the basis of reflection; and in the Application stage, they transfer those concepts to new and more complex situations (Kolb & Kolb, 2021). Grounded in the constructivist paradigms of Piaget and Vygotsky, the ARKA model ensures that knowledge is not merely received but actively and meaningfully constructed through social interaction and deep reflection (Ali, 2023; Kolb & Kolb, 2021). The cyclic structure of ARKA distinguishes it from linear instructional sequences by requiring students to revisit and refine their understanding at each successive stage, thereby embedding metacognition into the fabric of daily instruction.

To maximize the effectiveness of the ARKA model in developing critical thinking, its integration with the Science, Technology, Engineering, and Mathematics (STEM) approach constitutes a theoretically defensible strategic step. STEM supplies authentic, interdisciplinary problem contexts that inherently demand higher-order cognitive activity (English, 2023; Li & Schoenfeld, 2019). Complex STEM problems function as a Zone of Proximal Development

(ZPD) within Vygotsky's cognitive framework—an ideal space of challenge that enables students to transcend their current thinking capacity with the support of structured scaffolding (Kolb & Kolb, 2021; Wahono et al., 2020). In this sense, STEM is not merely a contextual backdrop but a cognitive catalyst that actively triggers the need for critical thinking. (English, 2023) argues that STEM problem solving, when carefully designed, engages students in iterative cycles of modeling, testing, and refining that parallel the cognitive processes of expert problem solvers. (Li & Schoenfeld, 2019) similarly caution against treating mathematics as a “given” within STEM and instead advocate problematizing mathematical ideas so that students experience genuine intellectual struggle.

Theoretically, the integration of ARKA with STEM enhances critical thinking through three mutually reinforcing mechanisms. First, the mechanism of active knowledge construction: ARKA is grounded in Kolb's Experiential Learning Theory, which asserts that knowledge is built—not passively received—through a cycle of concrete experience, reflective observation, abstract conceptualization, and active experimentation (Kolb & Kolb, 2021). When this cycle is filled with authentic STEM contexts, students are driven to interpret data, analyze inter-conceptual relationships, and evaluate the validity of solutions—three of Facione's six indicators that are particularly difficult to develop through conventional instruction. Second, the mechanism of productive cognitive dissonance: ill-structured STEM problems that lack a single correct answer create cognitive imbalance that propels students toward evidence-based inference and accountable explanation (Schukajlow et al., 2023; Thornhill-Miller et al., 2023). The Reflection and Conceptualization stages of ARKA supply structured space for resolving this dissonance through discussion, conclusion-drawing, and concept formalization. Third, the mechanism of metacognitive self-regulation: the Reflection stage of ARKA explicitly requires students to monitor and evaluate their own thinking processes, which is the essence of Facione's self-regulation indicator. Taken together, these three mechanisms form a theoretical hypothesis, not an established finding, that ARKA-STEM integration can engage all six Facione indicators within a learning cycle. The present study does not test this hypothesis directly.

Operationally, the linkage among ARKA stages, STEM contexts, and Facione indicators is hypothesized as follows. In the Activity stage, students confront STEM problems rooted in science and technology that demand interpretation (reading and making sense of contextual data) and analysis (identifying variables and mathematical inter-conceptual relationships). In the Reflection stage, group discussion of STEM exploration results activates evaluation (judging the logic of strategies employed) and self-regulation (reflecting on the thinking processes already undertaken). In the Conceptualization stage, concept formation from reflection supports inference (drawing conclusions from discovered patterns) and explanation (constructing systematic, verifiable conceptual accounts). In the Application stage, the transfer of concepts to new and more complex STEM problems integrates all six critical-thinking indicators simultaneously, because students must interpret novel situations, analyze conceptual relevance, evaluate solution feasibility, infer subsequent steps, explain solution arguments, and reflect on the quality of the completed solution (Ilyasa et al., 2024; Kolb & Kolb, 2021). This stage-to-indicator mapping is a design hypothesis; it was not independently tested here. What distinguishes ARKA from other structured, reflective STEM instruction is the required position of an explicit Reflection stage between the concrete task and formal Conceptualization, followed by a transfer task; in many PBL and PjBL implementations reflection is optional or placed at the end of a project.

This theoretical foundation is reinforced by a body of empirical evidence. (Rahmawati et al., 2022) reported, in a meta-analysis of studies on the STEM approach, an effect size of 1.074 on students' mathematical critical thinking (very large category). This finding is in line with the systematic review of (Azizah et al., 2025), which indicates that STEM-integrated learning contributes positively to the development of students' critical thinking. (Setyawati et al., 2022)

found that combining STEM with experience-based learning models significantly enhances mathematical critical thinking through active and collaborative engagement. (Schukajlow et al., 2023) further showed that real-world problem solving in interdisciplinary contexts significantly increases students' cognitive engagement and intrinsic motivation (Wahono et al., 2020, synthesizing 54 studies across Asia, reported an overall STEM effect size of 0.69 (95% CI [0.58, 0.81], moderate category) on student learning outcomes. (Khotimah, 2023) likewise demonstrated the effectiveness of STEM-based learning in improving critical thinking. Collectively, these findings provide supportive but largely indirect background for combining experiential cycles with STEM contexts; none of these studies evaluated the ARKA syntax.

A targeted search of Google Scholar, Scopus, and SINTA (keywords: “ARKA model,” “STEM mathematics,” “critical thinking Facione”; conducted through early 2026) identified few publications on the ARKA model, most of them in Islamic Religious Education and aimed at character strengthening (Ilyasa et al., 2024). Within that search we did not find an empirical study of ARKA in senior-high-school mathematics; this statement is limited to the search described and does not exclude work indexed elsewhere or published in other languages. STEM research in mathematics has largely examined PBL or PjBL (Ayu Cahyanti et al., 2024; Mukhibin & Herman, 2025), and in the sources we reviewed critical thinking was rarely assessed across all six Facione indicators. These observations indicate an under-explored area, not a demonstrated absence of prior work.

The contribution of this study is deliberately modest and is stated in terms of what the design can and cannot show. Conceptually, it examines one specific instructional syntax in which an explicit Reflection stage, placed before formal Conceptualization, is embedded in STEM problem contexts. This yields a testable expectation: if the reflection-before-formalization sequence matters, gains should appear in the indicators most closely tied to judging one's own strategies (evaluation and self-regulation). The present design cannot test this expectation, because ARKA and STEM were delivered together, compared only with conventional instruction, and each indicator was measured by a single item. Methodologically, all six Facione indicators are represented in the test, but at one item per indicator. Contextually, the study reports data from a non-urban school in Central Kalimantan, a setting with limited published evidence, although a single school cannot represent the region. The study is therefore a preliminary, descriptive comparison of two classes.

Given the gap between the demands of Kurikulum Merdeka and classroom realities—where 68% of SMAN 3 Sampit students had not reached the KKTP in mathematics and 53.13% were classified as low in critical thinking—pedagogical alternatives warrant systematic examination. This study examines the implementation of the ARKA model integrated with STEM in Grade XII mathematics at SMAN 3 Sampit and compares the critical-thinking scores of two intact classes. The research questions are: (1) How is the ARKA model integrated with STEM implemented in mathematics learning at SMAN 3 Sampit? (2) How large is the change in mathematical critical thinking scores in the ARKA-STEM class from pretest to posttest? (3) How do posttest scores of the ARKA-STEM class and the conventionally taught class compare after adjustment for pretest scores, and does this contrast vary with initial score? Because one class represents each condition, answers to questions 2 and 3 describe these two classes and are not estimates of a general treatment effect.

METHOD

Research Design

This study used a quasi-experimental pretest–posttest design with two intact classes. Random assignment of individual students was not feasible, so the class was the unit of assignment: from the four parallel Grade XII classes, two were drawn by lottery, the first drawn serving as the ARKA-STEM class and the second as the comparison class. The lottery and the pretest were both held on 20 July 2026, and the lottery was completed before the pretest was administered. Consequently, exactly one class represented each condition, and treatment is

fully confounded with class membership, including class composition, schedule, and group dynamics. The ARKA-STEM class received instruction based on ARKA integrated with STEM; the comparison class received conventional direct instruction (lecture, worked examples, and routine exercises). Both classes studied Circle Equations and Their Elements (Analytic Geometry) in three 90-minute meetings. The design supports a transparent description of how these two classes compared; it does not support inference about a treatment effect that generalizes beyond them (Creswell & Creswell, 2022).

Population and Sample

The population consisted of all Grade XII students of SMAN 3 Sampit in the 2026/2027 academic year (four parallel classes, 132 students). All four classes were eligible. Two were drawn by lottery on 20 July 2026, before the pretest was administered: XII Science-3 as the ARKA-STEM class ($n = 33$) and XII Science-1 as the comparison class ($n = 33$). A further parallel class, outside the two, was used to pilot the instrument. The topic matched the school's Alur Tujuan Pembelajaran (ATP) for Grade XII. The ARKA-STEM class began with a higher pretest mean (11.06 vs. 9.82; difference = 1.24 points, 95% CI $[-0.60, 3.09]$, about 5.2 points on the 0–100 scale; $t(64) = 1.347$, $p = .183$; Cohen's $d = 0.33$). A non-significant pretest test with 33 students per class would not demonstrate equivalence in any case, so pretest was entered as a covariate and baseline imbalance is treated as a limitation. The first author, who is also the mathematics teacher, taught both classes.

Research Variables and Operational Definitions

The independent variable was the instructional model (ARKA-STEM versus conventional direct instruction). The dependent variable was mathematical critical thinking ability, operationally defined as students' performance on a constructed-response test measuring the six indicators of (Facione, 2020): interpretation (understanding and clarifying mathematical data), analysis (identifying relationships among concepts and arguments), evaluation (judging the quality and logic of strategies), inference (drawing conclusions based on evidence), explanation (presenting and justifying reasoning systematically), and self-regulation (monitoring and evaluating one's own thinking processes). Each indicator was assessed by one essay item scored on a 0–4 analytic rubric, producing a total raw score ranging from 0 to 24. For reporting convenience, scores were also converted to a 0–100 scale by multiplying the raw score by $100/24$. The 0–4 rubric for each item specified performance descriptors ranging from “no evidence of the skill” (0) to “complete, accurate, and well-justified demonstration of the skill” (4), enabling consistent scoring across students and time points.

Instruments

Three instruments were employed. First, a six-item constructed-response critical-thinking test was developed by the researcher from Facione's indicators and the Circle Equations content, with one item per indicator (Table 1; the full items, test blueprint, and the 0–4 analytic rubric with descriptors for every score level are given in Supplementary Table S1). Five items were posed in purely mathematical form and one (Item 6) in a STEM context, a signal tower with a coverage radius. Each item was scored 0–4, giving a total of 0–24, and the same test served as pretest and posttest. Content validity was evaluated by two experts—a mathematics education lecturer at Universitas Muhammadiyah Sampit (Dr. Anwar Muttaqien, S.Pd., M.Pd.) and a senior mathematics teacher at SMAN 2 Sampit (Amalia Ayunani, S.Pd., M.Pd., using Aiken's V (acceptance threshold $V \geq 0.60$, set as the operational criterion of the study); with one rater this gives limited evidence of expert agreement. A pilot with 33 Grade XII students from a parallel class gave Cronbach's $\alpha = .929$. In the study data, α was .817 (pretest) and .781 (posttest) for all students; by class (ARKA-STEM/comparison) it was .784/.841 at pretest and .531/.709 at posttest. The low posttest α in the ARKA-STEM class reflects a narrow score range (12–23) and a mean inter-item Spearman correlation of only .182,

so individual total scores are measured with limited precision there, and with one item per indicator no subscale reliability can be estimated. Second, an observation checklist recorded the proportion of ARKA-STEM syntax elements enacted in each meeting; the checklist contained 22 indicators (14 on teacher activities across the four ARKA stages and 8 on student activities), each scored 1–4 by two observers. The meeting score is the mean of the two observers' scores (maximum 88), and fidelity is that score divided by 88. The indicators and scores are listed in Supplementary Table S3. No comparable checklist was applied to the comparison class. Third, five-point Likert questionnaires measured student and teacher responses to the model (syntax clarity, STEM integration, critical-thinking development, and practicality); items were reviewed for face validity by two education specialists.

Table 1. Test Blueprint: One Item per Facione Indicator

Item	Indicator	Item focus (Circle Equations)	Max. score
1	Interpretation	Determine the elements of a circle from a given equation	4
2	Analysis	Analyze the relationship between center and radius	4
3	Evaluation	Judge the correctness of a worked solution for a circle equation	4
4	Inference	Draw a conclusion about the position of a point relative to a circle	4
5	Explanation	Explain the steps for determining a circle equation	4

Note. Scores of 0 and 4 correspond to “no evidence of the skill” and “complete, accurate, and well-justified demonstration,” respectively; descriptors for scores 1–3 are given in Supplementary Table S1.

Data Collection Procedures

Data collection proceeded in three phases. In the pre-intervention phase, both classes completed the pretest under identical conditions (same time allocation, room arrangement, and invigilation protocol). In the intervention phase, the ARKA-STEM class received three ARKA-STEM lessons and the comparison class received conventional instruction; the three ARKA-STEM meetings were held on 24 August, 27 August, and 3 September 2026, and two observers rated implementation fidelity during each of them (22 indicators; Supplementary Table S3). A summary of the STEM task used in each meeting is given in Supplementary Table S2. In the post-intervention phase, both classes completed the same test as a posttest under conditions identical to the pretest, and the ARKA-STEM students completed the student questionnaire and the teacher (the first author) completed the teacher questionnaire. Testing sessions were timed (90 minutes) and supervised. All answer sheets were scored by the researcher using the analytic rubric; scoring was not blinded to class membership, and no second rater was involved, so inter-rater agreement could not be computed.

Ethical Considerations and Data Availability

The study was conducted with the written permission of the principal of SMAN 3 Sampit, supported by a research permit letter (No. E.5.d/624/DPPs-UMM/VI/2026, 26 June 2026) issued by the Directorate of Postgraduate Programs, Universitas Muhammadiyah Malang, and addressed to the school. Students and their parents or guardians were informed about the purpose and procedures of the research and gave written informed consent, and students took part voluntarily in the tests and questionnaires. Student identities were anonymized with respondent codes (E-01 to E-33 for the ARKA-STEM class and K-01 to K-33 for the comparison class), and the analysis files contain only these codes. No ethics-committee review was conducted for this study; research permission was obtained as described above. Anonymized item-level data (Supplementary Dataset S1) and the analysis script (Supplementary File S5) are provided as Supplementary Material.

Data Analysis Techniques

Descriptive statistics, including means, standard deviations, minimum and maximum scores, and frequency distributions, were used to summarize pretest, posttest, and normalized gain (N-Gain) scores. Within-class changes were examined using paired-samples t-tests, while

Wilcoxon signed-rank tests were applied to indicator-level changes. Between-class differences in posttest scores were examined using independent-samples t-tests and Mann–Whitney U tests. Analysis of covariance (ANCOVA) was employed to estimate differences in posttest scores after adjusting for pretest performance. Statistical assumptions were assessed using the Shapiro–Wilk test for normality, Levene’s test for homogeneity of variance, and an examination of regression slope homogeneity. Because the assumption of homogeneous regression slopes was not satisfied, an ANCOVA model incorporating the interaction between group and pretest scores was also examined, with conditional group differences reported at selected pretest values.

Additional diagnostic and sensitivity analyses were conducted to assess the stability of the estimated differences. Influential observations were identified using Cook’s distance, while students in the comparison class with unusually low pretest scores were examined separately. Sensitivity analyses included the exclusion of selected observations, heteroscedasticity-consistent HC3 robust standard errors, and 5,000 bootstrap replications. These procedures were undertaken after inspection of the primary results and were therefore treated as exploratory stability checks rather than prespecified analyses. Residual diagnostics included assessments of normality, variance homogeneity, heteroscedasticity, studentized residuals, influential observations, and leverage. These analyses evaluated the robustness of the observed estimates without resolving potential violations of statistical assumptions or limitations inherent in the research design.

Effect sizes were quantified using Cohen’s *d* (Cohen, 1988), Hedges’ *g*, and partial eta-squared, with 95% confidence intervals reported where available. Cohen’s *d* was calculated using pooled standard deviations, while effect sizes for between-class differences in score gains were based on the pooled standard deviation of individual gain scores. N-Gain was calculated separately for each student following Hake (1998) and subsequently summarized at the class level. N-Gain values were classified as high (greater than 0.70), moderate (0.30–0.70), or low (below 0.30). Negative values were retained as indicators of declining performance, and values equal to 0.70 were classified as moderate. Differences in indicator-level gains between classes were examined using Mann–Whitney U tests with Holm adjustment for multiple comparisons, accompanied by rank-biserial correlations as measures of effect size. Classroom observation and questionnaire data were analyzed descriptively.

Four descriptive benchmarks were established before analysis to evaluate the direction and magnitude of the observed results: statistically significant pretest-to-posttest improvement in the ARKA-STEM class, a nominally significant pretest-adjusted difference favoring that class, a mean N-Gain of at least 0.30, and a standardized between-class difference of at least 0.50, corresponding to a medium effect size (Cohen, 1988). These benchmarks were used to characterize the observed differences rather than establish instructional effectiveness. Because the study involved only one intact class per condition, instructional treatment was inseparable from class membership, and between-class variability could not be independently estimated. Consequently, student-level statistical tests, associated *p*-values, confidence intervals, and effect sizes were interpreted as descriptive measures of the contrasts between the two participating classes rather than evidence of a generalizable causal treatment effect.

RESULTS AND DISCUSSION

Implementation of the ARKA-STEM Model

Classroom observation indicated that the ARKA-STEM model was implementable and that fidelity improved across the three meetings. Fidelity is the obtained score divided by the maximum possible score; the three values correspond to mean observer scores of 62.5, 77.5, and 85.5 of 88 points (22 indicators scored 1–4 by two observers; Supplementary Table S3). It was recorded only in the ARKA-STEM class, so implementation quality cannot be compared between conditions. In Meeting 1, implementation reached 71.02% (good category); in Meeting 2 it rose to 88.07% (very good); and in Meeting 3 it attained 97.16% (very good). The

overall mean fidelity of 85.42% falls in the very good category. The progressive increase is consistent with the gradual familiarization of both teacher and students with the four-stage cycle of Activity, Reflection, Conceptualization, and Application. Early difficulties centered on managing time within the Reflection stage and on ensuring that STEM contexts remained mathematically rich rather than merely decorative. By Meeting 3 these difficulties had largely been resolved, and the observers recorded near-complete enactment of all planned syntax elements.

Table 2. Implementation Fidelity of the ARKA-STEM Model by Meeting

Meeting	Mean observer score (max 88)	Fidelity (%)	Category
1	62.5	71.02	Good
2	77.5	88.07	Very Good
3	85.5	97.16	Very Good
Mean	75.17	85.42	Very Good

Note. Categories follow school-based criteria: $\geq 85\%$ = Very Good; 70–84% = Good; 55–69% = Fair; $< 55\%$ = Poor. Percentages are obtained score divided by maximum score per meeting; fidelity was not recorded in the comparison class.

Student questionnaire responses (31 of 33 students; two absences) averaged 86.97% and were uniformly categorized as very positive. Teacher responses averaged 88.10% and ranged from positive to very positive across the dimensions of ARKA syntax, STEM integration, critical-thinking development, and practicality. These acceptance data suggest that the model is perceived as pedagogically coherent and classroom-feasible, although the possibility of ceiling effects and social-desirability bias in the questionnaires cannot be ruled out. Open-ended comments from students highlighted appreciation for the concrete STEM contexts and for the opportunity to discuss and revise their thinking during the Reflection stage. The teacher questionnaire was completed by the first author, who taught both classes, so it is a self-assessment of the researcher's own implementation and is descriptive only; actual implementation was assessed with the observers' checklist (Table 2).

Improvement of Critical Thinking in the Experimental Class

On the 0–24 scale, the experimental class improved from a pretest mean of 11.06 (SD = 3.60) to a posttest mean of 17.36 (SD = 3.29). The mean gain of 6.30 points was statistically significant, $t(32) = 18.939$, $p < .001$, 95% CI [5.62, 6.98]. Every student in the experimental class showed a positive gain (minimum gain = +2.00). Mean N-Gain was 0.510 (SD = 0.176), which falls in the moderate category according to Hake's (1998) criteria. The distribution of N-Gain categories revealed that 5 students (15.2%) reached the high category, 25 students (75.8%) the moderate category, and only 3 students (9.1%) remained in the low category. These results describe a substantial and fairly uniform pre–post increase in this class; part of it may reflect repeated use of the same test (see below).

Table 3. Descriptive Statistics and N-Gain for Critical Thinking Scores (0–24 Scale)

Group	N	Pretest M (SD)	Posttest M (SD)	Gain	N-Gain M (SD)	Category
Experimental	33	11.06 (3.60)	17.36 (3.29)	6.30	0.510 (0.176)	Moderate
Control	33	9.82 (3.89)	11.27 (4.30)	1.45	0.112 (0.128)	Low

Note. N-Gain categories: high ($g > 0.70$), moderate ($0.30 \leq g \leq 0.70$), low ($g < 0.30$) based on Hake (1998). Total-score ranges (0–24): ARKA-STEM pretest 4–18, posttest 12–23; comparison pretest 1–17, posttest 1–19. There were no missing data on the pretest or posttest for the 66 analyzed students. A student with an N-Gain of exactly 0.70 is classified as moderate. Score histograms are in Supplementary Figure S1.

The comparison class also showed a statistically significant but small gain, from a pretest mean of 9.82 to a posttest mean of 11.27 (gain = 1.45), $t(32) = 5.09$, $p < .001$. Mean N-Gain was 0.112 (low category); 31 of 33 comparison students (93.9%) were in the low category and four students declined. Because the same test was used on both occasions, part of this gain probably reflects test familiarity. The size of such an effect may differ between classes, particularly if the ARKA-STEM lessons rehearsed content close to the test items, and the design cannot separate these components. The contrast in N-Gain distributions is therefore

described as a feature of these two classes: three-quarters of ARKA-STEM students had moderate gains and about one in seven had high gains, whereas most comparison students remained in the low category.

Between-Group Differences After Controlling for Pretest

Because the ARKA-STEM class began with a higher pretest mean, pretest-adjusted comparisons are reported. Under a common slope, the adjusted means were 16.80 (SE = 0.31) and 11.84 (SE = 0.31), a difference of 4.96 points (95% CI [4.08, 5.84]) on the 0–24 scale, $F(1, 63) = 127.17$, $\eta^2p = .669$ (20.67 points on the 0–100 scale, 95% CI [17.01, 24.34]). This model assumes equal slopes, but the group $\times$ pretest interaction was significant, $F(1, 62) = 4.594$, $p = .036$: the slope of posttest on pretest was steeper in the comparison class (1.02) than in the ARKA-STEM class (0.78). The common-slope estimate is therefore a weighted summary and not one constant difference, and conditional differences from the interaction model are reported in Table 5. Removing three low-pretest comparison students or three influential cases changed the adjusted difference little (20.2–20.7 points on the 0–100 scale), and HC3 standard errors and bootstrap replications gave similar estimates (Table 4). These checks show that the estimate does not hinge on a few observations. They do not validate the common-slope model, and the change in the slope-heterogeneity p -value from .036 to .132 after deleting three students does not show that the heterogeneity is spurious. The exclusion rules were applied after the primary results were inspected and were not prespecified. None of these analyses addresses the one-class-per-condition structure.

Table 4. ANCOVA Results and Stability Checks (0–100 Scale)

Model	N	Adj. Diff.	95% CI	F (df)	η^2p	Slope-homogeneity p
– Full sample (common slope)	66	20.67	17.01–24.34	127.17 (1,63)	0.669	.036
– Without three low-pretest comparison students	63	20.27	16.55–23.98	118.88 (1,60)	0.665	.132
– Without three influential cases	63	20.23	16.83–23.63	141.59 (1,60)	0.702	.164
– HC3 robust SE	66	20.67	SE = 1.84	–	–	–
– Bootstrap (5,000)	66	20.66	17.12–24.25	–	–	–

Note. Adj. Diff. = adjusted mean difference (ARKA-STEM – comparison). Group effect $p < .001$ in all models. Stability checks do not correct the violation of the slope assumption or the design.

The interaction model (Table 5) shows that the conditional difference was larger at low pretest scores (28.96 points at a pretest of 10 on the 0–100 scale) and smaller at higher scores (14.20 at 70). This could reflect greater benefit for initially weaker students, but also ceiling constraints on higher scorers or regression toward the mean, which cannot be separated here. The estimate at 80 extrapolates beyond observed pretest values (maximum 75.00 and 70.83) and is not interpreted. A mixed ANOVA gave a group $\times$ time interaction of $F(1, 64) = 122.19$, $p < .001$, $\eta^2p = .656$; it describes differential change but, like the other student-level tests, treats 66 students as independent and cannot isolate treatment from class. Diagnostics for the interaction model were acceptable: residuals did not depart from normality (Shapiro–Wilk $p = .277$), residual variance did not differ by class (Levene $p = .722$; Breusch–Pagan $p = .977$), the largest studentized residual was 2.78, and Cook’s distance exceeded $4/n$ for one observation (maximum 0.130; in the common-slope model three observations exceeded $4/n$).

Table 5. Conditional Group Differences by Pretest Score (Interaction Model, 0–100 Scale)

Pretest	Difference (ARKA-STEM – comparison)	95% CI
10	28.96	20.45–37.47
30	24.04	19.29–28.79
50	19.12	15.28–22.97

Pretest	Difference (ARKA-STEM – comparison)	95% CI
70	14.20	7.20–21.21
80 (extrapolated)	11.75	2.69–20.80

Effect Sizes

Effect sizes are reported as descriptive summaries of the contrast between the two classes. Computed from the item-level data, Cohen's d for the posttest difference was 1.59 (95% CI [1.04, 2.14]; mean difference = 6.09; pooled SD = 3.83) and Hedges' g = 1.57. Cohen's d for the pre–post gain difference was 2.72 (95% CI [2.05, 3.39]; pooled SD of individual gains = 1.78) and for N-Gain 2.59 (95% CI [1.94, 3.24]); partial eta-squared from the ANCOVA was .669. Values of this size are unusual in educational research and should be read cautiously. They come from one pair of classes, a comparison condition without equivalent exploratory activity, a researcher-developed test aligned with the experimental lessons, unblinded scoring, and a posttest α of .531 in the ARKA-STEM class. The confidence intervals are wide and do not include class-level uncertainty, which cannot be estimated with one class per condition.

Indicator-Level Analysis

Because each indicator is measured by a single 0–4 item and posttest α in the ARKA-STEM class was .531, indicator-level results are exploratory. Posttest scores exceeded pretest scores on every indicator in the ARKA-STEM class (Wilcoxon signed-rank tests, all $p < .001$). Gain differences between classes (Mann–Whitney U with Holm correction across six tests) were significant for all indicators: Holm-adjusted $p < .001$ for interpretation, analysis, inference, and self-regulation, and $p = .010$ for evaluation and explanation (Table 6). Raw gain differences were largest for self-regulation (1.03; $r = .64$), interpretation (1.00; $r = .66$), and inference (0.97; $r = .72$), and smallest for evaluation (0.52; $r = .37$) and explanation (0.48; $r = .34$). The group $\times$ indicator interaction on gain scores was not significant, $F(5, 320) = 2.22$, $p = .052$, so the data do not show that the advantage differed across indicators.

Table 6. Mean Scores and Gain Differences per Critical Thinking Indicator (0–4 Scale)

Indicator	Experimental		Control		Gain diff. (95% CI)	r	Holm p
	pretest	posttest	pretest	posttest			
Interpretation	1.64	2.88	1.55	1.79	1.00 (0.67–1.33)	0.66	< .001
Analysis	1.79	2.82	1.61	1.79	0.85 (0.50–1.20)	0.57	< .001
Evaluation	1.97	2.79	1.45	1.76	0.52 (0.18–0.85)	0.37	.010
Inference	1.94	3.03	1.70	1.82	0.97 (0.69–1.25)	0.72	< .001
Explanation	1.82	2.82	1.70	2.21	0.48 (0.12–0.84)	0.34	.010
Self-Regulation	1.91	3.03	1.82	1.91	1.03 (0.67–1.39)	0.64	< .001

Note. Gain diff. = ARKA-STEM gain minus comparison gain (raw, not covariate-adjusted). r = rank-biserial correlation. Holm p adjusts across six indicators. One item per indicator; exploratory. 95% CIs for the gain difference are Welch t intervals.

Interestingly, the observed pattern of gains does not map neatly onto the a-priori theoretical assignment of indicators to ARKA stages. Evaluation and self-regulation were both hypothesized to be activated primarily in the Reflection stage, yet their gain differences differed by a factor of two; inference and explanation were both hypothesized to be activated in Conceptualization, yet they likewise differed substantially. This discrepancy suggests that the stage-to-indicator mapping should be treated as a conceptual hypothesis rather than an empirically verified mechanism. Process data (for example, scores on stage-specific learning sheets) would be required to test the mapping rigorously. The relatively smaller gains on evaluation and explanation may also reflect the greater linguistic and argumentative demands of these skills, which require extended written justification that some students found challenging within the time constraints of the test. With one item per indicator, differences among indicators may also reflect item difficulty and wording rather than any stage-to-indicator mechanism.

Proportion of Students Reaching the Mastery Threshold

The school's mastery criterion (70 on the 0–100 scale) is used only as a descriptive reference point. It was set for school achievement tests and has not been validated for this research-specific critical-thinking instrument, so the counts below do not indicate curricular mastery. In the ARKA-STEM class, 3/33 students (9.1%) scored ≥ 70 at pretest and 20/33 (60.6%) at posttest; in the comparison class the figures were 1/33 (3.0%) and 2/33 (6.1%). Reuse of identical items and closer alignment between the test and the ARKA-STEM lessons may inflate these proportions.

Discussion

All four descriptive benchmarks were met: the ARKA-STEM class improved significantly, the pretest-adjusted contrast favored it, its mean N-Gain reached the moderate category, and standardized differences were large. In this pair of classes, scores in the ARKA-STEM class therefore rose much more than in the comparison class. These results do not show that the model caused the difference, because the model, the class, and its group dynamics are inseparable, and because the measurement limitations discussed below could by themselves produce part of the gap.

The pattern is compatible with Experiential Learning Theory (Kolb & Kolb, 2021) and Facione's (2020) framework, but compatibility is not a test of mechanism: ARKA and STEM were not evaluated separately, and no stage-level process data were collected. The observed contrasts exceed the average STEM effects reported in meta-analyses (Rahmawati et al., 2022 : $ES = 1.074$ for mathematical critical thinking; Wahono et al., 2020: $ES = 0.69$ for learning outcomes in general), but the comparison is not like-for-like. Those estimates pool many designs, durations, outcome measures, and comparison conditions, whereas here the comparison was conventional direct instruction, the test was researcher-developed and aligned with the STEM activities, and the intervention was short; each of these tends to enlarge differences. The numerical gap should not be read as evidence that ARKA-STEM outperforms other STEM approaches.

Comparison with individual studies that examined STEM and critical thinking yields a more nuanced picture. (Setyawati et al., 2022), in a systematic review of STEM-PjBL, reported consistent positive effects on mathematical critical thinking but did not provide pooled effect sizes that permit direct numerical comparison. (Azizah et al., 2025) similarly concluded that STEM-integrated learning positively affects critical thinking, again without a single summary effect size. Khotimah (2023) found STEM-based learning effective for critical thinking, and (Rahmawati et al., 2022) reported significant effects of STEM on both critical and creative mathematical thinking. The present study adds a preliminary, descriptive data point from one school: an experiential-learning syntax combined with STEM contexts accompanied larger score gains in a three-meeting unit, with all six Facione indicators represented in the test (one item each).

Several limitations constrain inference. First and most important, one intact class represented each condition, so treatment is confounded with class membership (teaching context, group composition, and group dynamics). Student-level analyses treat students as independent although they share instruction, so nominal p-values and confidence intervals overstate precision, and the ANCOVA, mixed ANOVA, and t-tests describe these two classes without supporting population-level causal statements. Second, the researcher taught both classes, developed the test, and scored all answers without blinding to class; no second rater was used, so expectancy and instrument-alignment bias are possible. Third, the pretest and posttest used identical items, so testing effects are present (visible in the comparison class) and may differ by class. Fourth, each indicator was measured by one item, content validity rested on one expert, and posttest α in the ARKA-STEM class was .531; total scores should be read with caution and indicator scores as exploratory. The lesson plans also did not list the content of two items (relative position of two circles in Item 2; the circle through three points in Item

5) among the subtopics taught, so those items may measure transfer rather than taught content, in both classes alike. Fifth, the ARKA-STEM class began with a higher pretest mean, and slope heterogeneity ($p = .036$) means that no single adjusted difference describes all students. Sixth, ARKA and STEM were delivered as one package, so their separate contributions cannot be identified. Seventh, fidelity was recorded only in the ARKA-STEM class and only at meeting level; no equivalent record exists for the comparison class. Eighth, no delayed posttest was given, so durability is unknown. Ninth, the 70-point mastery criterion is not validated for this instrument.

Taken together, the descriptive results are consistent with larger score gains in the ARKA-STEM class, and this pattern is stable across parametric, non-parametric, and robust specifications. Stability across specifications, however, reflects the same two classes' data analyzed in different ways and not independent replications of the treatment, so it does not strengthen the causal interpretation. The findings are best read as preliminary, hypothesis-generating evidence that the ARKA-STEM syntax is feasible in a Grade XII classroom (fidelity rose from 71.02% to 97.16%, suggesting that a brief familiarization period matters) and that it merits replication.

From a theoretical standpoint, the results are compatible with the view that structured experiential cycles embedded in authentic problem contexts can support the Facione skills, but they do not test it. The imperfect alignment between predicted and observed indicator-level gains cautions against over-interpreting the a priori stage-to-indicator mapping, which should be treated as a conceptual hypothesis. Future research that collects process data at each ARKA stage is needed to refine or revise these linkages. From a practical standpoint, the study offers teachers a concrete four-stage framework that can be piloted on other topics, provided adequate preparation time and suitable problem materials are available.

CONCLUSION

Three conclusions are drawn, each limited to the two classes studied. First, the ARKA-STEM model could be implemented in a Grade XII mathematics class at SMAN 3 Sampit: observed fidelity rose from 71.02% to 97.16% (mean 85.42%), and student (86.97%) responses were very positive, with the teacher response (88.10%) being a self-assessment by the first author, although fidelity was not recorded in the comparison class and questionnaire responses may reflect social desirability. Second, critical-thinking scores in the ARKA-STEM class rose from 11.06 to 17.36 on the 0–24 scale (gain = 6.30, 95% CI [5.62, 6.98]; mean N-Gain = 0.510), with increases on all six indicators, while the comparison class rose from 9.82 to 11.27. Third, after adjustment for pretest, the ARKA-STEM class scored 4.96 points higher (95% CI [4.08, 5.84]), with a larger conditional difference at lower pretest scores and a standardized posttest difference of $d = 1.59$ (95% CI [1.04, 2.14]). Because one class represented each condition, these differences cannot be attributed to the model alone, and the study does not establish its effectiveness. Replication with multiple classes per condition, blinded scoring with inter-rater reliability, and a validated instrument is needed before any claim of effectiveness.

RECOMMENDATION

For mathematics teachers, the ARKA-STEM model may be adopted gradually, beginning with a single topic and allowing adequate time for familiarization with the four-stage cycle. Special attention should be given to the design of reflective prompts that encourage self-regulation and to the selection of STEM contexts that are both mathematically rich and culturally relevant to students. Teachers should also anticipate that the Reflection stage may require more time management in early meetings and should plan accordingly. For schools, support in the form of teaching materials, collaborative planning time, and professional development focused on STEM integration would facilitate wider adoption; pilot implementation in a limited number of classes is advisable before school-wide scaling. Provision of simple technology tools (for example, dynamic geometry software or data-

collection sensors) can enrich the STEM component without imposing prohibitive resource demands.

For future researchers, several methodological improvements are recommended: (a) employment of multilevel designs with multiple classes (or schools) per condition so that class-level variance can be modeled and generalizability enhanced; (b) comparison of ARKA-STEM with other innovative models such as PBL or PjBL rather than solely with conventional instruction, in order to locate the model's relative effectiveness within the broader landscape of active-learning pedagogies; (c) use of parallel test forms and blind scoring with inter-rater reliability reporting; (d) inclusion of at least three items per critical-thinking indicator to improve subscale reliability; (e) collection of process data (for example, stage-specific learning-sheet scores or think-aloud protocols) to test the hypothesized mapping of ARKA stages onto Facione indicators; (f) administration of delayed posttests to examine retention of gains over weeks or months; and (g) experimental designs that disentangle the relative contributions of the ARKA structure and the STEM context, for instance through a factorial design with ARKA-only, STEM-only, ARKA-STEM, and conventional conditions. For policy makers, the present findings should be regarded as preliminary evidence from a single site; multi-site replication across diverse geographic and socioeconomic contexts is necessary before the model is considered for large-scale curriculum or professional-development initiatives.

ACKNOWLEDGMENTS

The author expresses his deepest gratitude to the mathematics teachers and students at SMAN 3 Sampit who have actively participated and worked together during the implementation of this research. The author also expresses his appreciation to the supervisors and colleagues for the guidance, constructive input, and support provided during the research process

FUNDING INFORMATION

This research did not receive external funding

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Darus Salam	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	
Dyah Worowirastrri Ekowati	✓	✓		✓		✓		✓	✓	✓	✓	✓		
Agus Tinus	✓		✓							✓		✓		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

ETHICAL APPROVAL

The study was conducted with the written permission of the principal of SMAN 3 Sampit, supported by a research permit letter (No. E.5.d/624/DPPs-UMM/VI/2026, 26 June 2026) issued by the Directorate of Postgraduate Programs, Universitas Muhammadiyah Malang, and addressed to the school. Students and their parents or guardians gave written informed consent, and students participated voluntarily. All student data were anonymized with respondent codes. No ethics-committee review was conducted for this study

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy considerations of research participants

REFERENCES

- Ali, C. A. (2023). Preservice teachers' experiences in experiential instruction in place value. *Indonesian Journal of Science and Mathematics Education*, 6(1), 11–25. <https://doi.org/10.24042/ijjsme.v6i1.10440>
- Ayu Cahyanti, N., Suyanto, S., Wantara, N., A. Begimbetova, G., Unayah, H., & Rasyid, M. (2024). A systematic review of STEM education implementation in Indonesian high schools: Opportunities, challenges, and policy recommendations. *Jurnal Pendidikan MIPA*, 25(3), 1428–1443. <https://doi.org/10.23960/jpmipa/v25i3.pp1428-1443>
- Azizah, F., Waluya, S. B., & Ardiansyah, A. S. (2025). Systematic review on critical thinking

- through STEM integrated learning. *International Journal of Education in Mathematics, Science and Technology*, 13(6), 1384–1398. <https://doi.org/10.46328/ijemst.5106>
- Cevikbas, M., Kaiser, G., & Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competence: State-of-the-art developments in conceptualizing, measuring, and fostering. *Educational Studies in Mathematics*, 109(2), 205–236. <https://doi.org/10.1007/s10649-021-10104-6>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Creswell, J. W., & Creswell, J. D. (2022). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications.
- Dinas Pendidikan Kabupaten Kotawaringin Timur. (2024). *Data pendidikan Kabupaten Kotawaringin Timur tahun 2024*.
- English, L. D. (2023). Advancing STEM education through problem solving. *ZDM Mathematics Education*. <https://doi.org/10.1007/s11858-023-01474-7>
- Facione, P. A. (2020). *Critical thinking: What it is and why it counts*. Insight Assessment. <https://www.insightassessment.com/CT-Resources/Teaching-For-and-About-Critical-Thinking/Critical-Thinking-What-It-Is-and-Why-It-Counts/Critical-Thinking-What-It-Is-and-Why-It-Counts-PDF>
- González-Pérez, L. I., & Ramírez-Montoya, M. S. (2022). Components of Education 4.0 in 21st century skills frameworks: Systematic review. *Sustainability*, 14(3), Article 1493. <https://doi.org/10.3390/su14031493>
- Greefrath, G., Siller, H.-S., Vorhölter, K., & Kaiser, G. (2022). Mathematical modelling and discrete mathematics: Opportunities for modern mathematics teaching. *ZDM – Mathematics Education*, 54(4), 865–879. <https://doi.org/10.1007/s11858-022-01339-5>
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. <https://doi.org/10.1119/1.18809>
- Ilyasa, F. F., Rahardja, M. N. A., Firmansyah, M. I., Faqihuddin, A., & Muflih, A. (2024). ARKA model implementation (Aktivitas, Refleksi, Konseptualisasi, Aplikasi) in Islamic education learning: Building 21st century competencies. *Al-Tadzkiyyah: Jurnal Pendidikan Islam*, 15(1), 179–195. <https://doi.org/10.24042/002024152157300>
- Kemdikbudristek. (2023). *Laporan hasil asesmen nasional 2022: Ringkasan eksekutif*. Pusat Asesmen dan Pembelajaran.
- Khotimah, R. P. (2023). STEM-based learning effectiveness in improving critical thinking. *Cogent Education*, 10. <https://doi.org/10.1080/2331186X.2023.2220233>
- Kolb, D. A., & Kolb, A. Y. (2021). Experiential learning theory: A dynamic, holistic approach to management learning, education and development. In *The SAGE handbook of management learning, education and development*. SAGE Publications.
- Li, Y., & Schoenfeld, A. H. (2019). Problematizing teaching and learning mathematics as “given” in STEM education. *International Journal of STEM Education*, 6(1), Article 44. <https://doi.org/10.1186/s40594-019-0197-9>
- Mukhibin, A., & Herman, T. (2025). How STEM education research conducted in mathematics learning in Indonesia? A systematic literature review. In *Technological advancements and innovations in future of education. Transactions on Computer Systems and Networks*. Springer. https://doi.org/10.1007/978-981-97-4931-7_31
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*.
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- Paul, R., & Elder, L. (2020). *The miniature guide to critical thinking concepts and tools* (8th ed.). Foundation for Critical Thinking.

- Rahmawati, L., Juandi, D., & Nurlaelah, E. (2022). Implementasi STEM dalam meningkatkan kemampuan berpikir kritis dan kreatif matematis. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, 12(1), 319–331. <https://doi.org/10.24127/ajpm.v12i1.6044>
- Schukajlow, S., Rakoczy, K., & Pekrun, R. (2023). Emotions and motivation in mathematics education: Where we are today and where we need to go. *ZDM – Mathematics Education*, 55(2), 249–267. <https://doi.org/10.1007/s11858-022-01463-2>
- Setyawati, R. D., Pramasdyahsari, A. S., Astutik, I. D., Nusuki, U., Aini, S. N., Arum, J. P., Widodo, W., Salmah, U., & Zuliah, N. (2022). Improving mathematical critical thinking skill through STEM-PjBL: A systematic literature review. *International Journal of Research in STEM Education*, 4(2), 1–17. <https://doi.org/10.33830/ijrse.v4i2.1141>
- Siahaan, E. Y. S., Muhammad, I., & Dasari, D. (2023). Trend of critical thinking skill researches in mathematics education in Scopus database across Indonesia: From research design to data analysis. *International Journal of Trends in Mathematics Education Research*, 6(2), 151–161. <https://doi.org/10.33122/ijtmr.v6i2.216>
- Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., & Bourgeois-Bougrine, S. (2023). Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills. *Journal of Intelligence*, 11(3), Article 54. <https://doi.org/10.3390/jintelligence11030054>
- Trisnani, N. (2024). Challenges of Indonesian elementary school mathematics teachers in integrating critical thinking into the classroom. *Journal on Mathematics Education*, 15(3), 905–924. <https://doi.org/10.22342/jme.v15i3.pp905-924>
- Wahono, B., Lin, P.-L., & Chang, C.-Y. (2020). Evidence of STEM enactment effectiveness in Asian student learning outcomes. *International Journal of STEM Education*, 7(1), Article 36. <https://doi.org/10.1186/s40594-020-00236-1>
- Wang, Q., & Abdullah, A. H. (2024). Enhancing students' critical thinking through mathematics in higher education: A systemic review. *SAGE Open*. <https://doi.org/10.1177/21582440241275651>
- Yaniawati, P., Kariadinata, R., Sari, N. M., Pramiasih, E. E., & Mariani, M. (2020). Integration of e-learning for mathematics on resource-based learning: Increasing mathematical critical thinking and self-confidence. *International Journal of Emerging Technologies in Learning*, 15(6), 60–78. <https://doi.org/10.3991/ijet.v15i06.11915>

Supplementary Table S1. Critical-thinking test, items, and analytic rubric

Topic: Circle Equations and Their Elements (Analytic Geometry). Six constructed-response items, one per Facione indicator; 90 minutes; 0–4 points per item; maximum total 24. English text is a translation of the Indonesian items administered (Instrument 1).

No.	Indicator (Bloom level)	Item (translation)
1	Interpretation (C4)	Given $x^2 + y^2 - 6x + 4y - 3 = 0$. Identify the center and radius, rewrite the equation in standard form $(x - a)^2 + (y - b)^2 = r^2$, and explain the geometric meaning of each element you found.
2	Analysis (C4)	Circle 1: $(x - 2)^2 + (y + 3)^2 = 25$; Circle 2: $x^2 + y^2 - 4x + 6y - 12 = 0$. Analyze the relationship between the circles: find the center and radius of each, compare their relative position (separate, externally tangent, internally tangent, intersecting, or one inside the other) by computing the distance between the centers, and give a complete mathematical justification.
3	Evaluation (C5)	A student determines the circle with center $(3, -1)$ passing through $(7, 2)$. Step 1: $r^2 = (7 - 3)^2 + (2 - (-1))^2 = 16 + 9 = 25$, so $r = 5$. Step 2: $(x + 3)^2 + (y - 1)^2 = 25$. Step 3: the circle has center $(3, -1)$ and radius 5. Evaluate the answer: is each step correct? If there are errors, locate them and write the correct solution with mathematical reasons.
4	Inference (C5)	Circle L: $x^2 + y^2 - 8x + 2y + 7 = 0$ and points A(4, 1), B(1, 3), C(2, -1). Without drawing, determine whether each point is inside, on, or outside L, draw a valid conclusion from your calculations, and explain the criterion used.
5	Explanation (C5)	Determine the equation of the circle through P(1, 3), Q(5, 3), and R(3, 7). Explain completely and systematically all steps from finding the center to writing the final equation, with the mathematical reason for each step.
6	Self-regulation (C6)	A signal tower stands at M(3, 4) with range 5 units; devices are at A(6, 8) and B(0, 1). (a) Model the coverage as a circle equation. (b) Decide whether A and B are within range. (c) Critically re-examine your whole solution: assumptions made, whether the method is most appropriate, other possible answers; write your reflection explicitly.

Analytic rubric (score 0–4 per item; translated from Instrument 1).

Item/indicator	0	1	2	3	4
1 Interpretation	No answer, or unrelated to interpreting a circle equation.	Basic understanding of the circle equation form but significant errors in completing the square or interpreting the elements.	Attempts completing the square but with calculation errors, so the center or radius is inaccurate.	Completes the square and identifies the elements correctly, but the geometric explanation is incomplete or there is one minor error.	Converts to standard form correctly, identifies center and radius accurately, and explains the geometric meaning of each element completely and systematically.
2 Analysis	No answer, or no evidence of analyzing the relationship	Identifies only some elements of one circle and cannot make a	Identifies one circle correctly and tries to compare, but	Identifies both circles correctly and computes the distance	Identifies both centers and radii accurately, computes the

Item/indicator	0	1	2	3	4
	between two circles.	valid comparison.	errs in the distance between centers or in the relationship.	between centers, but the conclusion is inaccurate or the reasoning incomplete.	distance correctly, and concludes the relationship with complete mathematical reasons.
3 Evaluation	No answer, or states the student's answer is correct without critical evaluation.	Only states the answer is wrong, without locating the error or giving a meaningful correction.	Knows there is an error and attempts a correction, but the correction is flawed or the reasoning inaccurate.	Locates the error accurately and gives the correct correction, but the reasoning is shallow or unsystematic.	Identifies the error in Step 2 (sign of the center) precisely, writes the correct equation $(x - 3)^2 + (y + 1)^2 = 25$, and gives a clear, logical explanation.
4 Inference	No answer, or no understanding of the position of a point relative to a circle.	Attempts substituting points but the results and conclusions are wrong.	Determines the position of only one point correctly, or all points with an inappropriate method or criterion.	Determines at least two points correctly with the right criterion, but the explanation or one conclusion is inaccurate.	Determines all three positions correctly by substitution, uses the correct criterion ($f(x, y) < 0, = 0$, or > 0), and draws valid, complete conclusions.
5 Explanation	No answer, or an explanation irrelevant to determining the circle equation.	Writes only initial steps or a small part of the process without reaching the required equation.	Attempts to explain the steps, but a conceptual or procedural error at one stage affects the final result.	Explains the steps correctly and reaches the right final equation, but the reason at one step is missing or shallow.	Explains all steps systematically and completely: (1) find the center (perpendicular bisectors or a system of equations), (2) compute the radius, (3) write the final equation, with a clear mathematical reason at each step.
6 Self-regulation	No answer, or no evidence of self-regulation and re-checking.	Only forms the circle model (part a) but fails part (b) or writes no reflection.	Solves (a) and (b) correctly but writes no critical reflection, or the reflection	Solves (a) and (b) correctly and writes a reflection, but it is incomplete	Forms the model correctly, determines the positions of both devices correctly, and

Item/indicator	0	1	2	3	4
			is irrelevant to the work done.	(e.g., only one aspect).	reflects critically and explicitly: states assumptions, evaluates the suitability of the method, and considers alternative answers.

Score conversion: $\text{score}/24 \times 100$; categories: 21–24 very high; 17–20 high; 13–16 moderate; 9–12 low; 0–8 very low.

Supplementary Table S2. Lessons in the two classes (three meetings of 2×45 min)

ARKA-STEM class: opening 10 min; Activity 15; Reflection 15; Conceptualization 20; Application 20; closing 10 (group work in fours, STEM worksheet LKPD-1 to LKPD-3, GeoGebra or grid paper). Comparison class: opening 10; teacher explanation 30; worked examples 20; question-and-answer 10; individual textbook exercises 15; closing 5.

Meeting / topic	STEM context and ARKA tasks (ARKA-STEM class)	Reflection prompts (ARKA-STEM class)	Comparison class
1. Circle and standard equation	Context: radar and GPS coverage, satellite dish. Activity: plot six points exactly 4 units from $P(2, 3)$ on grid paper and verify with the distance formula. Conceptualization: derive $(x - a)^2 + (y - b)^2 = r^2$ from the distance definition. Application: graded tasks, ending with a WiFi-router coverage problem (router at $(0, 0)$, range 15 m; is a room at $(9, 12)$ covered?).	How do you define a circle mathematically? What happens if the radius is doubled? Why can radar and GPS use the geometry of the circle?	Teacher explains the definition and standard equation; worked examples on the board; individual textbook exercises.
2. General equation	Context: satellite orbit and gear design. Activity: expand $(x - 2)^2 + (y - 3)^2 = 25$ into $x^2 + y^2 + Dx + Ey + F = 0$. Conceptualization: derive conversion formulas for a, b, r from D, E, F . Application: ends with a power-plant safe-zone problem ($x^2 + y^2 - 10x - 8y + 16 = 0$; is a settlement at $(12, 10)$ inside?).	What pattern links a, b, r to D, E, F ? Why might an engineer prefer the general form in design software?	Teacher explains the general form and conversion; worked examples; individual exercises.
3. Position of points and lines	Context: radar range zone. Activity: classify six points against $x^2 + y^2 = 100$ by comparing $x^2 + y^2$ with r^2 . Conceptualization: criteria from substituting a point into the equation. Application: ends with a drone-in-signal-range problem ($(x - 5)^2 + (y - 3)^2 = 100$, drone at $(13, 9)$).	Which criterion tells whether a point is inside, outside, or on the circle? How can it extend to a line and a circle?	Teacher explains the criteria; worked examples; individual exercises.

Supplementary Table S3. Implementation Fidelity Checklist and Observed Scores (22 indicators)

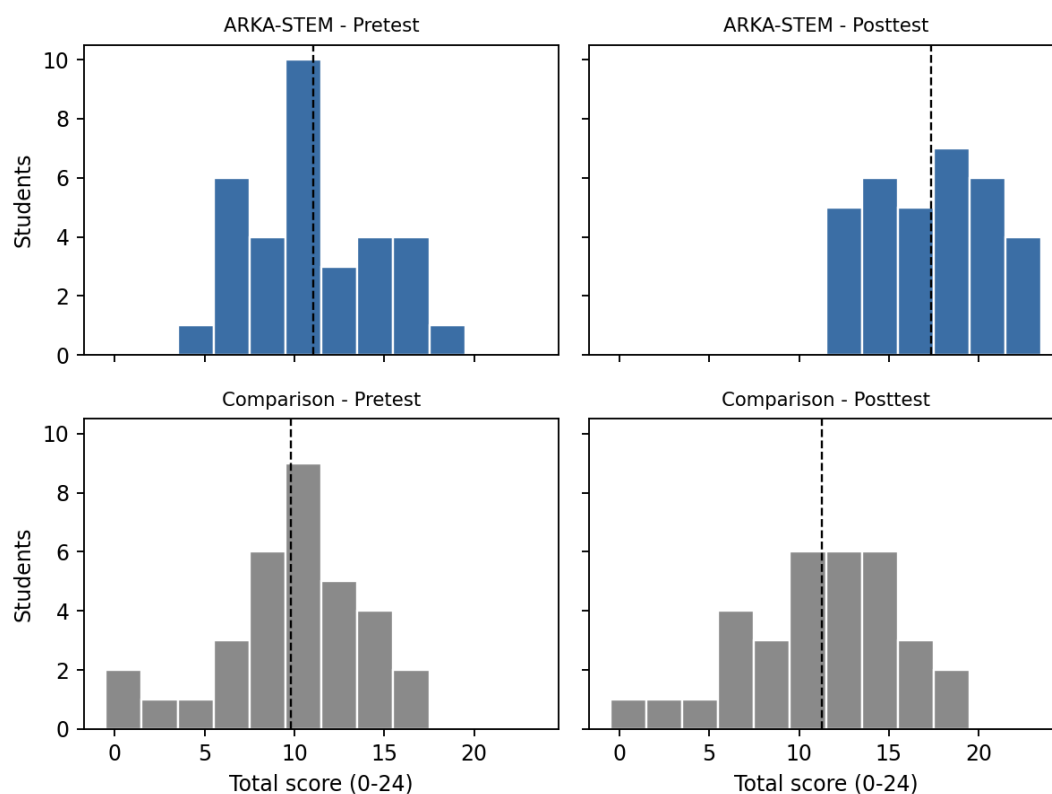
Two observers rated every indicator on a 1–4 scale (1 = poor, 4 = very good) in each of the three ARKA-STEM meetings. Cells show the mean of the two observers. Meeting score = sum of the 22 indicator means (maximum 88); fidelity (%) = meeting score / 88 × 100.

Code	Indicator	Meeting 1	Meeting 2	Meeting 3
Activity Stage				
A1	The teacher presents contextual STEM-based problems (e.g., radar or GPS coverage).	3	4	4
A2	The teacher facilitates group exploration using STEM-integrated student worksheets.	3	4	4
A3	The teacher guides students in collecting data and identifying patterns related to circle equations.	3	3	4
A4	The teacher provides scaffolding during group discussions.	3	4	4
Reflection Stage				
R1	The teacher encourages students to express difficulties or ideas encountered during exploration.	2.5	4	4
R2	The teacher poses guiding questions to help students connect and reflect on their exploratory experiences.	2.5	3	4
R3	The teacher facilitates comparisons of findings and solutions across groups.	2	3.5	4
Conceptualization Stage				
K1	The teacher guides students in formulating the general equation of a circle.	3	4	4
K2	The teacher reinforces formal mathematical concepts derived from students' findings.	3	3	4
K3	The teacher addresses misconceptions that emerge during discussions.	3	3	3.5
K4	The teacher summarizes the key concepts related to circle equations.	2.5	4	4
Application Stage				
Ap1	The teacher provides contextual STEM-based mathematical modeling exercises.	3	4	4
Ap2	The teacher monitors students as they solve application problems independently or in pairs.	3	3	4
Ap3	The teacher provides feedback and reviews students' problem-solving results.	3	4	4
Student Activities				
S1	Students observe and interpret the contextual STEM-based problems presented.	3	4	4
S2	Students actively participate in group discussions to complete the student worksheets.	3	3	4
S3	Students collect data and identify patterns of mathematical relationships.	3	3.5	4

Code	Indicator	Meeting 1	Meeting 2	Meeting 3
S4	Students communicate their ideas or difficulties during the Reflection stage.	2.5	3	4
S5	Students participate in conceptually formulating mathematical equations.	3	3.5	4
S6	Students record key concepts and explanations related to circle equations.	3	3	4
S7	Students demonstrate the ability to solve contextual STEM-based application problems.	3	4	3
S8	Students review and verify the accuracy of their problem-solving procedures.	2.5	3	3
	Total Score (Maximum = 88)	62.5	77.5	85.5
	Implementation Fidelity (%)	71.02	88.07	97.16

Exact agreement between the two observers on 86.4% of the 66 indicator ratings. Mean fidelity = 85.42% (Very Good). Categories: $\geq 85\%$ Very Good; 70–84% Good; 55–69% Fair; $< 55\%$ Poor. No comparable checklist was used in the comparison class.

Supplementary Figure S1. Score distributions by class and occasion (dashed line = mean)



Total scores (0–24), $n = 33$ per class. Generated by File S5 from Dataset S1.