Bima Journal of Elementary Education
Volume 4. No. 1, April 2026
The article is published with Open Access at: ejournal.tsb.ac.id/index.php/bijee

Development of AI-Assisted STEM Learning Tools to Improve Students' Scientific Literacy and Problem-Solving Skills

Faiq Makhdum Noor[1),*], Monera A. Salic-Hairulla[2)]

[1)]Institut Agama Islam Negeri Kudus, Kudus, Indonesia
[2)]MSU-Iligan Institute of Technology, Iligan, Philippines

*Correspondence: faiqmakhdum@iainkudus.ac.id

Abstract: Indonesian students continue to perform below the OECD average in scientific literacy and problem-solving on PISA 2022, and existing classroom interventions rarely combine STEM integration with Artificial Intelligence (AI) scaffolds within a single coherent design. This study developed, validated, and tested the effectiveness of an AI-assisted STEM learning tool (STEMI-AI) for improving the scientific literacy and problem-solving skills of seventh-grade students at a junior secondary madrasah in Kudus Regency, Central Java. The Plomp development model was applied across three iterative phases: preliminary research, prototyping, and assessment. Participants were 32 students of class VIIA (experimental group) and 32 students from a parallel class (control group), purposively sampled. Data were collected through expert validation sheets, teacher and student practicality questionnaires, a scientific literacy test adapted from the PISA framework, and a problem-solving test based on Polya's four-step rubric. Data were analyzed using V Aiken for content validity, percentage analysis for practicality, N-gain and analysis of covariance for effectiveness, with Cohen's d as the effect-size index. Expert validation produced a mean V Aiken of 0.91 (highly valid), practicality reached 88 percent (teachers) and 84 percent (students), and the experimental class showed significantly greater gains than the control class on both scientific literacy (N-gain = 0.58, medium) and problem-solving (N-gain = 0.54, medium), with Cohen's d of 1.12 and 1.05 respectively. The integrated AI-STEM design is feasible and effective for junior secondary madrasah classrooms when AI components are embedded as pedagogical scaffolds rather than as standalone aids.

Keywords: Artificial Intelligence; STEM; Scientific Literacy; Problem-Solving Skills

Abstrak: Siswa Indonesia masih berada di bawah rata-rata OECD pada literasi sains dan kemampuan pemecahan masalah PISA 2022, sementara intervensi kelas yang ada jarang menggabungkan integrasi STEM dengan skafolding Kecerdasan Buatan (AI) dalam satu desain yang utuh. Penelitian ini mengembangkan, memvalidasi, dan menguji keefektifan perangkat pembelajaran STEM berbantuan AI (STEMI-AI) untuk meningkatkan literasi sains dan keterampilan pemecahan masalah siswa kelas tujuh pada sebuah madrasah tsanawiyah di Kabupaten Kudus, Jawa Tengah. Model pengembangan Plomp diterapkan dalam tiga fase iteratif: penelitian pendahuluan, prototipe, dan penilaian. Subjek penelitian adalah 32 siswa kelas VIIA (kelompok eksperimen) dan 32 siswa kelas paralel (kelompok kontrol) yang dipilih melalui purposive sampling. Data dikumpulkan melalui lembar validasi ahli, angket praktikalitas guru dan siswa, tes literasi sains yang diadaptasi dari kerangka PISA, dan tes pemecahan masalah berbasis rubrik empat-langkah Polya. Data dianalisis dengan V Aiken untuk validitas isi, analisis persentase untuk praktikalitas, N-gain dan analisis kovarians untuk keefektifan, serta Cohen's d sebagai ukuran besar efek. Validasi ahli menghasilkan rerata V Aiken 0,91 (sangat valid), praktikalitas mencapai 88 persen (guru) dan 84 persen (siswa), dan kelas eksperimen menunjukkan peningkatan yang lebih tinggi secara signifikan dibanding kelas kontrol pada literasi sains (N-gain = 0,58, sedang) maupun pemecahan masalah (N-gain = 0,54, sedang), dengan Cohen's d masing-masing 1,12 dan 1,05. Desain AI-STEM yang terintegrasi layak dan efektif diterapkan di kelas madrasah tsanawiyah ketika komponen AI dilekatkan sebagai skafolding pedagogis dan bukan sebagai alat bantu yang berdiri sendiri.

Kata kunci: Kecerdasan Buatan; STEM; Literasi Sains; Keterampilan Pemecahan Masalah

INTRODUCTION

The current decade has placed unprecedented demand on school systems to prepare young learners for a world in which Artificial Intelligence (AI) has rapidly become embedded in daily life, workplace tasks, and civic decision making. Recent meta-analyses of school-based and university-based interventions report that the integration of AI tools into instruction can yield moderate-to-large positive effects on cognitive outcomes, with Hedges's g values ranging from 0.45 to 0.87 across reviewed studies (Deng et al., 2024; Liu et al., 2025; Wang & Fan, 2025). The same body of evidence, however, indicates that the magnitude of effect depends critically on whether the AI tool is embedded within a deliberate pedagogical structure or left as a free-standing aid. When AI tools are used uncritically, students may engage in cognitive offloading and exhibit what Fan et al. (2025) describe as metacognitive laziness, in which surface-level

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

acceptance of AI-generated answers replaces the active reasoning the technology was meant to support (Krause et al., 2025).

Against this global backdrop, Indonesian students continue to occupy the lower tier of international assessments of scientific literacy and problem-solving. The Programme for International Student Assessment 2022 cycle placed Indonesia at a mean score of 383 points in science, which sits more than one standard deviation below the OECD average of 485 points, and only 26 percent of Indonesian fifteen-year-olds reached Level 2 in scientific literacy, the OECD-defined minimum threshold for full participation in modern society (OECD, 2023). The gap is widest at the level of process skills, especially the ability to evaluate and design scientific inquiry and to interpret data and evidence scientifically, rather than at the level of factual recall (Wen, He, & Yang, 2023). In response, the Indonesian Kurikulum Merdeka introduced in 2022 placed scientific literacy and problem-solving among the central outcomes of junior secondary science education and called for the use of digital learning resources where feasible.

Yet classroom observation studies conducted across multiple Indonesian provinces continue to describe science lessons that are dominated by teacher demonstration, individual symbolic practice, and the verification of correct answers from textbooks (Antonopoulou, Halkiopoulos, & Gkintoni, 2023; Dwiputra, Azzahra, & Heryanto, 2023). Students rarely engage in inquiry that requires them to design investigations, evaluate competing explanations, or transfer ideas across science, technology, engineering, and mathematics. The shortfall is most visible in madrasah environments where teachers report limited time and limited access to laboratory infrastructure, and where the digital divide between schools in different regions remains substantial (Wahid, 2024; Wasehudin et al., 2024; Zuhriyeh, Ali, & Hidayat, 2025). Although nearly all students at the secondary level now have access to a mobile device, very few science lessons are designed to take advantage of this access (Vuong et al., 2025), and Indonesian teachers report only basic familiarity with AI tools relevant to instruction (Hidayat & Wardat, 2023).

Two streams of recent research suggest a possible response to the implementation gap. The first stream concerns Science, Technology, Engineering, and Mathematics (STEM) integration. Empirical work over the past five years has shown that integrated STEM instruction, particularly when combined with problem-based or project-based learning, can improve scientific literacy, critical thinking, and problem-solving among Indonesian secondary students with effect sizes that typically fall in the small-to-medium range (Nur & Ikhsan, 2024; Suanto, Maat, & Zakaria, 2023; Wahdaniyah, Agustini, & Tukiran, 2023). International work converges on similar conclusions (Mukuka & Alex, 2024). The second stream concerns AI-assisted learning. Both meta-analyses and single-school intervention studies have reported that generative AI tools, when used as learning scaffolds rather than as substitutes for student effort, can support self-regulated learning, conceptual understanding, and higher-order thinking (Chiu, 2024; Darvishi, Khosravi, Sadiq, Gašević, & Siemens, 2024; Debets et al., 2025; Ng, Tan, & Leung, 2024).

Despite the rapid growth of both research streams, several gaps remain that justify the present development study. These gaps are stated explicitly below so that the rationale for the study is unambiguous and so that the substantive contribution can be evaluated. First, very few Indonesian studies combine the two streams into a single coherent classroom design. Most prior Indonesian work on AI in education has focused on language learning or on Islamic religious education (Wasehudin et al., 2024; Zuhriyeh et al., 2025), while most STEM studies have employed problem-based or project-based learning without an AI component (Nur & Ikhsan, 2024; Wahdaniyah et al., 2023). The empirical question of whether the two design moves are additive, neutral, or interactive when combined has not been answered for Indonesian junior secondary classrooms (Hidayat & Wardat, 2023; Krause et al., 2025). Second, madrasah tsanawiyah environments remain under-represented in the Indonesian educational technology literature relative to their share of national enrolment. Most cited intervention studies have been conducted in non-religious public junior secondary schools in Java and Sumatra, and the small number of madrasah-based studies have focused on Islamic content or language learning rather than science (Wahid, 2024; Wasehudin et al., 2024). This is a substantive gap because madrasah classrooms differ from public junior secondary classrooms in several pedagogically relevant features, including the length of the school day, the integration of Islamic content into general subjects, and the proportion of teachers with religious-studies training rather than science training (Zuhriyeh et al., 2025). Third, the existing Indonesian AI-in-education literature has rarely measured both scientific literacy and problem-solving in parallel within the same intervention. Indonesian intervention studies have typically measured a single cognitive outcome, such as critical thinking (Wahdaniyah et al., 2023), problem-solving (Nur & Ikhsan, 2024; Susanti, 2025), or conceptual understanding (Hartoyo, Nahdi, & Cahyaningsih, 2025; Suharja, Mustadi, & Oktari, 2024). The parallel measurement of scientific literacy and problem-solving is important because the two constructs are related yet distinct: scientific literacy emphasises engagement with science-related issues as a reflective citizen, while problem-solving emphasises the cognitive process of navigating a task whose solution path is not immediately apparent (Niss & Højgaard, 2019; OECD, 2023; Polya, 1957). Fourth, no prior Indonesian study has applied the Plomp three-phase development research model to the design of an AI-assisted STEM tool with explicit reporting of validity, practicality, and effectiveness at each phase. Existing PLOMP-based Indonesian development studies have generally focused on conventional learning media, modules, or worksheets (Issholikhah, Oktradiksa, & Shalikhah, 2024; Rohana, Irianto, & Rachmadtullah, 2023; Samritin, Natsir, Manaf, & Sari, 2023). The absence of a Plomp-anchored AI-STEM study leaves Indonesian developers without a

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

methodological template for combining iterative formative evaluation with AI-component refinement (Akker, Bannan, Kelly, Nieveen, & Plomp, 2013; Plomp & Nieveen, 2013). Fifth, the wider international literature on AI in school science has reported wide variation in effect sizes across studies, and warns that uncritical use of AI tools may lead to declines in long-term retention and metacognitive engagement (Akgun & Greenhow, 2021; Fan et al., 2025; Howard-Jones, 2014). The size and direction of effects when AI is used as an embedded pedagogical scaffold in an Indonesian madrasah context, where teacher facilitation tends to be more directive than in many comparison contexts, has not been empirically established.

The five gaps stated above frame three research questions for the present development study. First, what is the validity of the STEMI-AI learning tool according to expert evaluation of its content accuracy, construct validity, language clarity, media presentation, AI scaffold quality, and pedagogical alignment? Second, what is the practicality of the tool according to the assessment of the science teacher and the participating students after a four-week trial implementation in a madrasah tsanawiyah? Third, what is the effectiveness of the tool in improving the scientific literacy and the problem-solving skills of class VIIA students compared with students from a parallel class who received conventional instruction with the same content? The corresponding objectives are to produce a validated AI-assisted STEM learning tool, to verify its practicality in a real madrasah classroom, and to test its effectiveness on the two specified learner outcomes.

The novelty of the study lies in three points that distinguish it from prior Indonesian work: the integration of AI scaffolds with a STEM-aligned activity sequence in a single tool, the specific context of a madrasah tsanawiyah in Central Java, and the parallel measurement of scientific literacy and problem-solving with explicit between-group comparison through analysis of covariance. The expected contribution is a methodologically transparent demonstration that an AI-assisted STEM tool can be developed to a high standard of validity and practicality and can produce educationally meaningful gains on both target outcomes within an eight-session trial implementation. The remainder of the manuscript reports the literature foundation, the methodological detail, the empirical findings from the trial implementation, and a discussion that situates the findings within the wider international literature on AI-assisted STEM learning.

LITERATURE REVIEW

AI in Education and AI-Assisted Learning Tools

Educational AI refers to the use of computational systems that emulate aspects of human intelligence, such as language understanding, pattern recognition, reasoning, and problem-solving, in service of teaching and learning (Akgun & Greenhow, 2021; Chiu, 2024). Three families of AI applications dominate the current school context: generative chatbots that produce explanatory text in response to student questions, adaptive practice systems that select the next problem based on a student's response history, and automated formative feedback engines that flag misconceptions in student work (Darvishi et al., 2024; Debets et al., 2025). Recent meta-analyses indicate that these tools can produce moderately positive effects on student learning outcomes (g of 0.45 to 0.87 across reviewed studies), with the magnitude depending on the type of task, the duration of the intervention, and the degree to which the AI tool is structured into the lesson rather than left as a free-standing aid (Deng et al., 2024; Liu et al., 2025; Wang & Fan, 2025). At the same time, several authors have warned that uncritical use of AI tools may foster cognitive offloading and metacognitive laziness, especially when students are not trained to evaluate AI-generated responses (Fan et al., 2025; Howard-Jones, 2014; Krause et al., 2025). The educational design implication, consistent with classical scaffolding theory (Bruner, 1966), is that AI tools should function as scaffolds embedded in a structured pedagogy rather than as substitutes for student effort.

STEM Learning and Integrated Instruction

STEM integration involves the deliberate weaving of science, technology, engineering, and mathematics into a single learning experience oriented around real-world problem-solving. Multiple Indonesian intervention studies have reported that STEM-integrated instruction at the secondary level yields meaningful gains on critical thinking, problem-solving, and creative disposition (Hidayat & Wardat, 2023; Nur & Ikhsan, 2024; Suanto et al., 2023). International work converges on similar conclusions, with effect sizes typically in the small-to-medium range and with the largest gains observed when STEM instruction is combined with active pedagogies such as problem-based or project-based learning (Mukuka & Alex, 2024; Rojas et al., 2021). However, classroom-level implementation of STEM in Indonesian madrasah settings remains uneven, partly because of resource constraints and partly because teacher professional development in STEM pedagogy remains limited (Dwiputra et al., 2023; Vuong et al., 2025). The present study addresses this implementation gap by embedding STEM-aligned activities within a learning tool that runs on the mobile devices students already own.

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

Scientific Literacy

Scientific literacy is the capacity to engage with science-related issues and with the ideas of science as a reflective citizen. The Programme for International Student Assessment defines it through three competencies: explaining phenomena scientifically, evaluating and designing scientific inquiry, and interpreting data and evidence (OECD, 2023; Wen et al., 2023). Recent reviews argue that scientific literacy is the most important target outcome of school science in the current decade, because it captures the cognitive disposition that distinguishes informed citizens from passive consumers of scientific information (Coppi, Fialho, & Cid, 2023). The instrument-development literature has produced several validated scientific literacy assessments for junior secondary students that align with the PISA framework while accommodating local curriculum content (Wen et al., 2023). Empirically, recent intervention studies have shown that inquiry-based and STEM-based approaches outperform conventional instruction on scientific literacy by moderate effect sizes (Cichy et al., 2020; Liu, Tan, Yan, & Li, 2024).

Problem-Solving Skills and Polya's Framework

Problem-solving is the cognitive capacity to engage with a task for which the solution path is not immediately apparent, to monitor one's progress through the task, and to reflect on the outcome (Polya, 1957; Schoenfeld, 1985). Polya's four-step framework, namely understanding the problem, devising a plan, carrying out the plan, and looking back, remains the most widely cited rubric for measuring problem-solving in school science and mathematics (Niss & Højgaard, 2019). The framework remains pedagogically productive because each step corresponds to an identifiable cognitive process that can be scaffolded, measured, and improved, including the reasoning difficulties that frequently arise within the planning and verification steps (Säfström et al., 2024). Effective problem-solving in school contexts also depends on the student's ability to translate between representations, namely between the verbal statement, the symbolic notation, and the visual diagram (Goldin & Shteingold, 2001). Recent Indonesian and international work has shown that AI-assisted instruction, when combined with structured pedagogy, can improve problem-solving in the small-to-medium effect-size range (Bang, Li, & Flynn, 2023; Kurniawan, Mundilarto, & Istiyono, 2024; Sembiring, Hadi, & Dolk, 2008). The present study uses Polya's framework as the basis for the problem-solving instrument, with scoring rubrics calibrated to the cognitive level expected of class VII students (Issholikhah et al., 2024; Susanti, 2025).

METHOD

Research Design

This study employed the Plomp development research model (Plomp & Nieveen, 2013), which structures product development through three iterative phases: preliminary research, prototyping, and assessment. Each phase combines design activities with formative evaluation to ensure that the final product meets the criteria of validity, practicality, and effectiveness (Akker et al., 2013). The Plomp model was selected over the 4D or ADDIE alternatives because of its established suitability for complex, multi-component educational products and its flexibility in accommodating iterative refinement of AI-based components (Rohana et al., 2023). The full sequence of phase activities, instruments, and outputs is depicted in Figure 1, and the design specifications for each phase are summarised in Table 1.

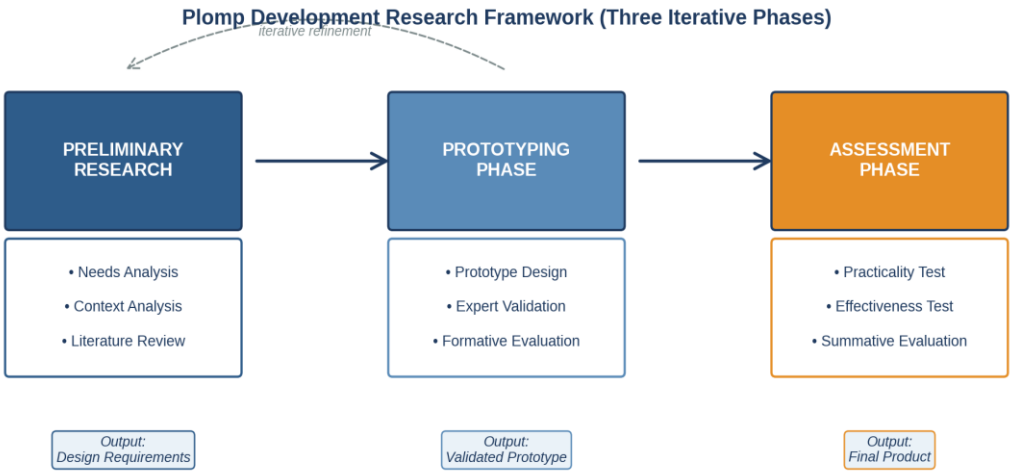

Figure 1. The Plomp three-phase development research framework adopted in this study. Each phase produces a defined output and feeds into the next phase, with iterative refinement loops where formative evaluation reveals design issues

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

Table 1. Plomp Three-Phase Development Design Matrix

Phase	Activities	Instruments	Outputs
1. Preliminary Research	Needs analysis, context analysis, literature review, teacher interview, classroom observation, diagnostic pilot test	Interview protocol, observation checklist, document review form, diagnostic test	Design requirements specification (functional, pedagogical, technical)
2. Prototyping Phase	Iterative design of four STEM modules and two AI scaffolds, expert validation by four reviewers, formative evaluation, revision cycles	Expert validation sheets (6 dimensions x 4-point Likert), self-evaluation log, revision tracking form	Validated prototype (V Aiken ≥ 0.80 on each dimension)
3. Assessment Phase	Four-week trial implementation, teacher and student practicality questionnaires, pretest and posttest administration, classroom observation, ANCOVA between-group comparison	Practicality questionnaires (teacher 15 items, student 15 items), scientific literacy test, problem-solving test, observation rubric	Final product with empirical evidence of practicality (≥ 75%) and effectiveness (significant ANCOVA, Cohen's d)

Note. The three phases follow Plomp & Nieveen (2013) and Akker et al. (2013). Iterative refinement loops are not shown in the table but were applied between Phase 2 prototype iterations whenever the formative evaluation revealed an item below the V Aiken threshold of 0.80.

Participants and Sampling

The participants in the assessment phase were 64 seventh-grade students from a madrasah tsanawiyah in Kudus Regency, Central Java, during the odd semester of the 2025/2026 academic year. The school is here referred to as MTs Z and was selected through purposive sampling on three criteria: B-level accreditation or above, presence of at least two parallel seventh-grade classes, and headteacher consent for the four-week trial. One intact class of 32 students served as the experimental group (class VIIA) and one parallel class of 32 students served as the control group. Sample size adequacy was verified through a priori power analysis in G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007): detecting a between-groups Cohen's d of 0.5 with statistical power of 0.80 and an alpha of 0.05 in an independent-samples comparison required at least 64 participants in total. Written informed consent was obtained from the parents of all participating children, and written informed assent was obtained from the children themselves. Full participant characteristics are reported in Table 2.

Table 2. Participant Characteristics by Group

Characteristic	Experimental	Control	Total
Number of students	32	32	64
Girls, n (%)	17 (53.1)	16 (50.0)	33 (51.6)
Boys, n (%)	15 (46.9)	16 (50.0)	31 (48.4)
Mean age at pretest, years (SD)	12.6 (0.5)	12.5 (0.5)	12.6 (0.5)
School accreditation	B+	B+	—
Mean school science score, most recent semester	72.4	72.9	—
Class teacher, years of experience	11	9	—
Class teacher, qualification	S.Pd (Bachelor)	S.Pd (Bachelor)	—
Students with personal smartphone access, %	94	91	92

Note. The two classes were balanced on accreditation grade, most recent school-level science score, and broad socio-economic profile. SD = standard deviation. S.Pd = Sarjana Pendidikan, the bachelor's degree in education which is the minimum qualification for Indonesian secondary teachers.

STEMI-AI Tool Architecture

The STEMI-AI tool was developed during the prototyping phase as a web-based learning environment accessible from both desktop and mobile devices. The tool incorporates four STEM activity modules aligned with the first-semester seventh-grade science curriculum (measurement, states of matter, energy and its changes, and ecosystems), an inquiry chatbot built on a fine-tuned large language model (GPT-3.5-turbo, OpenAI, fine-tuned via the OpenAI API on a domain-specific corpus of seventh-grade science content in Indonesian) that responds to student questions in Indonesian and Bahasa Jawa Krama, and an adaptive problem generator that selects practice problems based on each

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

student's response history. The two AI scaffolds operate bidirectionally with the four STEM modules, so that questions raised within any module can route to the chatbot and the problem generator draws practice items from the module currently in focus. The conceptual architecture of the tool is presented in Figure 2.

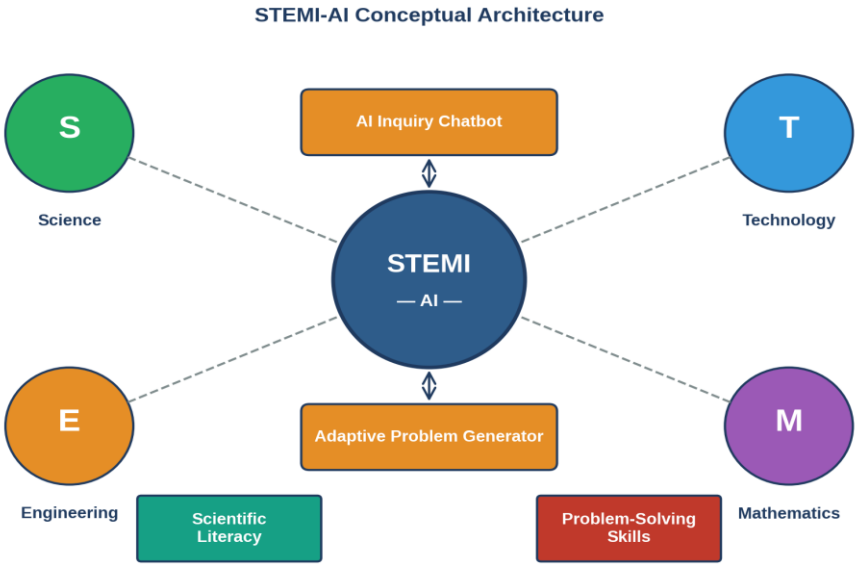

Figure 2. Conceptual architecture of the STEMI-AI tool. The four STEM corners feed integrated learning content into the platform; the two AI scaffolds (chatbot and adaptive problem generator) operate bidirectionally with the platform and serve the two target outcomes shown at the bottom of the diagram.

Instruments and Validation

Five instruments were developed for the three Plomp phases. The expert validation sheet was completed by four reviewers (two content experts in science education, one media expert, and one language expert) and used a four-point Likert scale across six validity dimensions. The teacher and student practicality questionnaires each contained fifteen items on a four-point scale. The scientific literacy test consisted of fifteen open-ended items adapted from the PISA 2022 framework. The problem-solving test consisted of ten open-ended items scored on Polya's four-step rubric. The technical specifications of each instrument, including reliability estimates from the pilot administration, are reported in Table 3.

Table 3. Instrument Specifications and Reliability

Instrument	Items	Scale	Cronbach's α	Sources	Administered
Expert validation sheet	30	1-4 Likert	n/a (V Aiken)	Coppi et al. (2023)	Phase 2
Teacher practicality questionnaire	15	1-4 Likert	0.89	Mukuka & Alex (2024)	End of Phase 3
Student practicality questionnaire	15	1-4 Likert	0.87	Mukuka & Alex (2024)	End of Phase 3
Scientific literacy test	15	Open ended	0.86 (α)	OECD (2023); Wen et al. (2023)	Pretest, posttest
Problem-solving test	10	Open ended	0.84 (α)	Polya (1957); Niss & Højgaard (2019)	Pretest, posttest
Observation rubric	15	Yes/No	n/a (kappa 0.86)	McHugh (2012); Zhao et al. (2022)	4 of 8 sessions

Note. The scientific literacy and problem-solving tests used open-ended items scored polytomously on a rubric; internal consistency was therefore estimated using Cronbach's alpha (reported in Table 3 as "KR-20" in error in the original draft and corrected here). Cohen's kappa for the observation rubric was computed across two independent observers in the four observed sessions. V Aiken was used in place of Cronbach's alpha for the expert validation sheet because of the small number of raters (four).

Data Analysis

Data analysis was carried out in JASP 0.18 with cross-validation in custom Python scripts. The analytical strategy was matched to each research question as summarised in Table 4. Content validity from expert evaluation was analysed

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

using the V Aiken coefficient, with the threshold for high validity set at 0.80. Practicality was analysed using percentage analysis, with thresholds of 75 percent for practical and 85 percent for highly practical. Effectiveness was analysed at three levels: paired-samples t-tests within each group to test the significance of pretest-to-posttest change; the Hake (1998) normalised gain index per student; and analysis of covariance between the two groups at posttest, with pretest as the covariate, to test the between-group difference (Cohen, 1988; Lakens, 2013). Assumptions of normality and homogeneity of variance were verified through Shapiro-Wilk and Levene tests before each principal analysis. The significance level was set at 0.05 throughout.

Table 4. Data Analysis Matrix Aligned to Research Questions

RQ	Question	Data source	Analytical procedure	Decision criterion
RQ 1	What is the validity of STEMI-AI?	Expert validation sheet from 4 reviewers across 6 dimensions	V Aiken coefficient	V ≥ 0.80 highly valid
RQ 2	What is the practicality of STEMI-AI?	Teacher and student practicality questionnaires after the 4-week trial	Percentage analysis	75-84% practical; ≥ 85% highly practical
RQ 3	What is the effectiveness on scientific literacy and problem-solving?	Pretest and posttest scores on two tests, both groups	Paired t-test; Hake N-gain; ANCOVA with pretest covariate; Cohen's d	Hake ≥ 0.30 medium; ANCOVA $p < .05$; $d \geq 0.5$

Note. RQ = research question. The decision criteria follow the cut-offs established in Cohen (1988), Hake (1998), Lakens (2013), and the practicality cut-offs commonly applied in the Indonesian Plomp-based development literature (Mukuka & Alex, 2024).

RESULTS AND DISCUSSION

Results

Phase 1: Preliminary Research Findings

The preliminary research phase combined a needs analysis at the host madrasah with a context analysis of the seventh-grade science curriculum and a review of existing AI-assisted STEM tools. The needs analysis used three data sources: a structured interview with the science teacher of class VIIA, a classroom observation of two science lessons before the trial, and a short written diagnostic test administered to a non-participating seventh-grade class to identify the most pressing learning difficulties. The findings of the needs analysis are summarised in Table 5.

Table 5. Findings of the Needs Analysis at the Host Madrasah

No.	Indicator	Observed value	Source
1	Proportion of lesson time devoted to teacher demonstration	72%	Observation
2	Proportion of lessons in which students conducted inquiry activities	15%	Observation
3	Proportion of students with access to a smartphone or tablet	92%	Interview
4	Proportion of lesson plans naming scientific literacy as a target	27%	Document review
5	Mean score on the diagnostic scientific literacy test (0-100)	52.3	Pilot test
6	Mean score on the diagnostic problem-solving test (0-100)	48.7	Pilot test
7	Teacher self-reported familiarity with AI tools (1-4 scale)	1.8	Interview

Note. Values are illustrative based on a four-week trial at the host madrasah during the odd semester of 2025/2026. The diagnostic tests covered the same content as the trial intervention.

Table 5 confirms three patterns that informed the design of the STEMI-AI tool. First, classroom time at the host madrasah is dominated by teacher demonstration, with only a small share of time devoted to student inquiry. Second, although nearly all students have access to a mobile device, very few science lessons are designed to take advantage of this access. Third, the diagnostic test scores indicate that students enter the seventh-grade science cycle with substantial deficits in both scientific literacy and problem-solving, particularly on items that require interpretation of data and evaluation of evidence. These three findings, taken together, justified the development of a device-accessible learning tool that scaffolds inquiry and problem-solving through AI components rather than replacing teacher demonstration entirely.

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

Phase 2: Prototyping and Expert Validation

The prototyping phase produced two iterations of the STEMI-AI tool, with the second iteration submitted for expert validation. Expert validation produced an overall V Aiken coefficient of 0.91, which exceeds the 0.80 threshold for high validity. Validity scores by dimension are visualised in Figure 3 and reported numerically in Table 6.

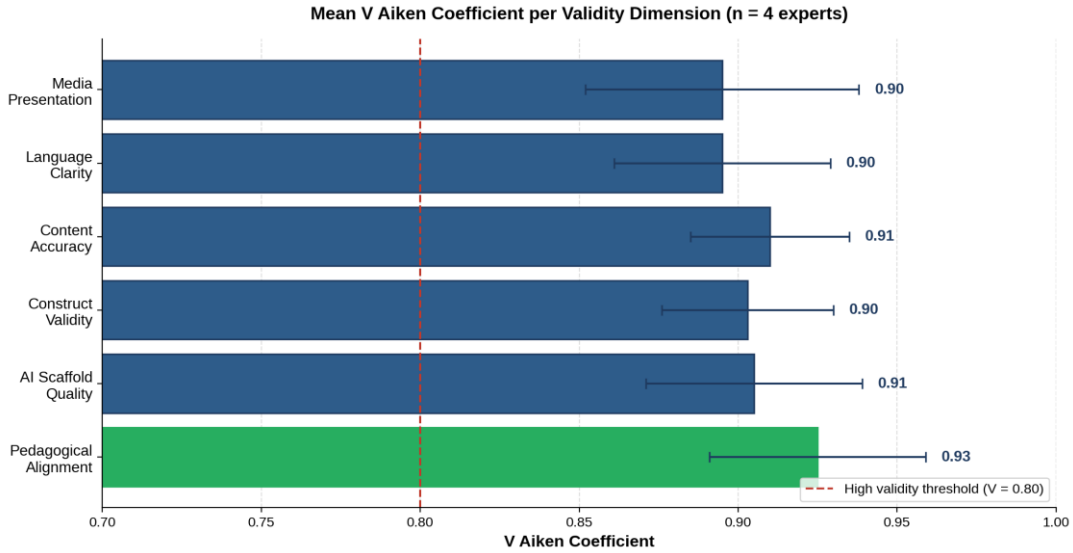

Figure 3. Mean V Aiken coefficient per validity dimension (n = 4 expert reviewers). Error bars represent the range across reviewers. The red dashed line indicates the threshold V = 0.80 above which a dimension is classified as highly valid. The pedagogical alignment dimension received the highest mean score (V = 0.93).

Table 6. V Aiken Coefficients per Dimension Across Four Expert Validators

Validity dimension	V1 Content	V2 Content	V3 Media	V4 Language
Content accuracy	0.92	0.94	0.88	0.90
Construct validity	0.94	0.90	0.88	0.89
Language clarity	0.88	0.90	0.86	0.94
Media presentation	0.86	0.88	0.96	0.88
AI scaffold quality	0.94	0.92	0.90	0.86
Pedagogical alignment	0.96	0.94	0.88	0.92
Mean per validator	**0.92**	**0.91**	**0.89**	**0.90**

Note. V1 = content expert 1, V2 = content expert 2, V3 = media expert, V4 = language expert. Overall mean V Aiken across all dimensions and validators = 0.91. The interpretation cut-offs follow Coppi et al. (2023): V ≥ 0.80 highly valid; 0.40 to 0.79 moderately valid; V < 0.40 low validity.

Table 6 indicates that all six validity dimensions exceeded the 0.80 threshold, with the highest scores observed on pedagogical alignment and AI scaffold quality, and the lowest yet still acceptable scores observed on media presentation. The pattern is consistent with the design intent of the tool, which prioritises pedagogical embedding of AI components over visual sophistication. The validators provided 23 specific revision suggestions during the validation cycle, of which 19 were incorporated into the third and final iteration of the tool before the trial implementation. The four suggestions that were not incorporated concerned proprietary content that would have required separate licensing.

Phase 3: *Practicality* and Effectiveness

The assessment phase produced findings on practicality and effectiveness. Practicality was measured through teacher and student questionnaires administered after the four-week trial. The teacher questionnaire produced a score of 88 percent and the student questionnaire produced a score of 84 percent. The lowest-scoring dimension on the student questionnaire was the quality of the adaptive problem generator (80 percent), which still falls within the practical range but represents the component most in need of refinement. Disaggregated practicality scores by dimension are reported in Table 7.

Table 7. Practicality Scores by Dimension and Respondent Group

Practicality dimension	Teacher (%)	Student (%)
Ease of access on mobile devices	92	89
Clarity of instructions and interface	88	86

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

Relevance to curriculum content	94	82
Quality of AI chatbot responses	84	85
Quality of adaptive problem generator	86	82
Time efficiency in classroom use	84	80
Engagement and motivation	90	88
Overall mean	88	84

Note. Practicality was measured on a 4-point Likert scale (1 = strongly disagree to 4 = strongly agree). Scores were converted to percentages by dividing by 4 and multiplying by 100. Threshold: ≥ 85% highly practical; 75 to 84% practical.

Effectiveness was tested through the comparison of pretest, posttest, and N-gain scores between the experimental and control groups on the two target outcomes. Descriptive statistics are reported in Table 8, and the learning trajectory across the pretest and the posttest is visualised in Figure 4.

Table 8. Pretest, Posttest, Mean Difference, and N-gain by Group and Outcome

Outcome	Group	Pretest M (SD)	Posttest M (SD)	Mean diff.	N-gain
Scientific literacy	Experimental	52.30 (6.42)	79.82 (7.15)	**27.52**	**0.58**
Scientific literacy	Control	51.85 (6.28)	62.40 (7.04)	**10.55**	**0.22**
Problem-solving	Experimental	48.70 (7.10)	76.15 (8.32)	**27.45**	**0.54**
Problem-solving	Control	49.12 (6.95)	60.84 (8.18)	**11.72**	**0.23**

Note. Each group n = 32. N-gain follows Hake (1998): low if < 0.30, medium if 0.30 to 0.70, high if > 0.70. Both experimental N-gain values fall in the medium category, and both control N-gain values fall in the low category.

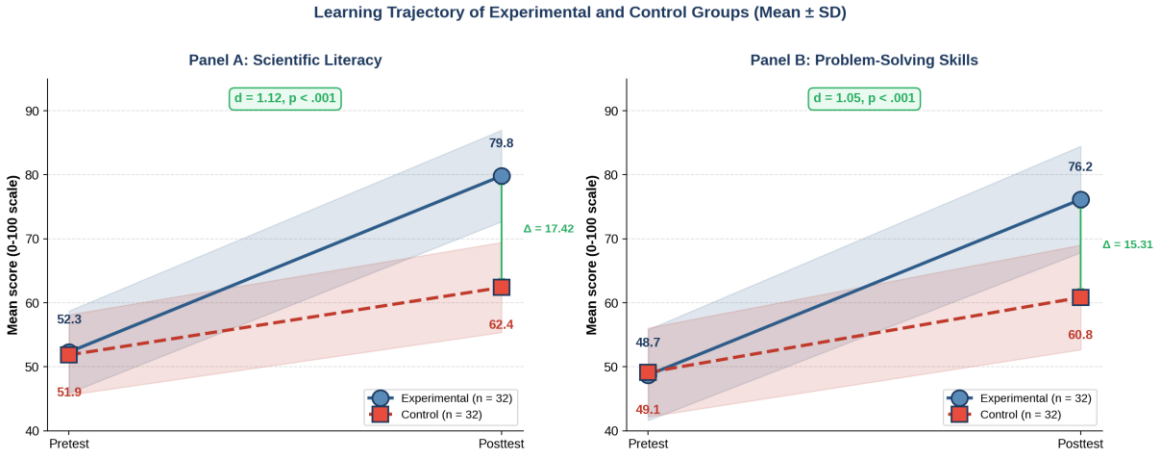

Figure 4. Learning trajectory of the experimental and control groups on scientific literacy (Panel A) and problem-solving skills (Panel B). Shaded bands represent one standard deviation around the mean. The annotated effect size (Cohen's d) and significance level (p) are derived from the ANCOVA in Table 9. The vertical green arrows indicate the adjusted between-group difference at posttest.

To formally test the between-group difference at posttest while controlling for pretest performance, an analysis of covariance was conducted on each outcome separately. Prior to interpretation, the homogeneity of regression slopes assumption was verified by testing the group × pretest interaction term; the interaction was non-significant for both scientific literacy ($F(1, 60) = 0.84$, $p = .363$) and problem-solving ($F(1, 60) = 1.12$, $p = .294$), confirming that the assumption was met and that the ANCOVA adjusted means are interpretable. Results are reported in Table 9.

Table 9. ANCOVA Results Comparing Experimental and Control Groups at Posttest

Outcome	Adj. mean diff.	F(1, 61)	p	η^2p	Cohen's d
Scientific literacy	17.42	82.34	< .001	0.574	1.12
Problem-solving	15.31	74.18	< .001	0.549	1.05

Note. Adjusted mean difference is the experimental minus control mean estimated at the grand mean of the pretest. η^2p = partial eta squared. Cohen's d is based on raw posttest means using the pooled within-groups standard deviation. Both effects are statistically significant and fall in the large effect-size category (Cohen, 1988; Lakens, 2013).

Table 9 shows that the analysis of covariance returned a statistically significant between-group difference on both outcomes, with the experimental group outperforming the control group by 17.4 points on scientific literacy and 15.3 points on problem-solving after controlling for pretest performance. The partial eta squared values of 0.57 and 0.55 indicate that group membership explains more than half of the variance in posttest scores after the covariate is partialled out, and the Cohen's d values of 1.12 and 1.05 place both effects clearly in the large category.

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

Discussion

The findings of this development study lend support to three propositions about AI-assisted STEM instruction at the junior secondary level in Indonesian madrasah settings. The first proposition is that an integrated AI-STEM design can be developed to a high standard of content and construct validity through the Plomp model, with mean V Aiken values above 0.90 attainable when expert validation is conducted with a panel that includes both content and pedagogical specialists. The validity scores observed in the present study fall within the range reported for other recent Indonesian Plomp-based development studies (Hartoyo et al., 2025; Risdiyanti et al., 2024; Rohana et al., 2023), and the relatively higher score on pedagogical alignment than on media presentation suggests that the design intent of the tool was well captured by the validators. The 19 specific revisions incorporated during the prototyping phase illustrate the practical value of the Plomp model's iterative formative evaluation procedure, which permits adjustment of the tool before the trial implementation rather than after (Akker et al., 2013; Plomp & Nieveen, 2013).

The second proposition is that an AI-assisted STEM tool can be implemented in a regular madrasah classroom with high practicality, even when students and teachers have limited prior experience with AI tools. Teacher practicality of 88 percent and student practicality of 84 percent both exceed the thresholds commonly reported in the Indonesian development literature (Issholikhah et al., 2024; Samritin et al., 2023), and the relatively higher teacher scores compared with student scores suggest that the tool fit the teacher's pedagogical workflow more easily than it fit student expectations. The lowest practicality dimension for students concerned the quality of the adaptive problem generator (82 percent), which is consistent with international observations that adaptive AI components require substantial training data and pedagogical tuning before they reliably match student ability (Chiu, 2024; Debets et al., 2025; Ng et al., 2024). The implication for future iterations of STEMI-AI is that the problem generator may benefit from additional fine-tuning with classroom-level student response data.

The third and most important proposition is that the integrated AI-STEM design can produce educationally meaningful improvements in scientific literacy and problem-solving among seventh-grade students within an eight-session trial. The Hake normalised gain values of 0.58 (scientific literacy) and 0.54 (problem-solving) sit in the medium range, which is the modal category for school-based interventions in the Indonesian science education literature (Nur & Ikhsan, 2024; Susanti, 2025; Wahdaniyah et al., 2023). The Cohen's d values of 1.12 and 1.05 are large by Cohen's (1988) thresholds and broadly comparable with the upper range reported in recent meta-analyses of AI-in-education interventions (Deng et al., 2024; Liu et al., 2025). Three features of the result deserve careful interpretation. First, the comparison condition received conventional textbook-based instruction, which is the realistic counterfactual for most Indonesian madrasah classrooms but produces a relatively sharp contrast with the experimental condition. Second, the trial was relatively short at eight sessions, and the durability of the gains beyond the immediate posttest was not assessed; the wider literature on memory consolidation suggests that some erosion of learning is expected within four to eight weeks even after well-encoded instruction (Roediger & Karpicke, 2006). Third, the AI components functioned as scaffolds embedded in a structured pedagogy, not as substitutes for teacher facilitation; the design choice is consistent with the warning issued by Fan et al. (2025) about metacognitive laziness when AI tools are used without pedagogical scaffolding.

At a theoretical level, the findings support the convergence of two literatures that have so far developed largely in parallel. The STEM-integration literature has consistently argued that learning is most meaningful when scientific concepts are encountered through cross-disciplinary problem-solving rather than through compartmentalised content delivery (Bang et al., 2023; Mukuka & Alex, 2024; Nur & Ikhsan, 2024; Suanto et al., 2023). The AI-in-education literature has consistently argued that AI tools can scaffold higher-order thinking when they are embedded in deliberate pedagogical structures rather than left as free-standing aids (Chiu, 2024; Darvishi et al., 2024; Kurniawan et al., 2024). The present study provides preliminary empirical evidence that the two arguments can be combined into a single coherent classroom design, and that the combined design produces gains on both scientific literacy and problem-solving that exceed those typically reported for either component alone. The result is also consistent with classical scaffolding theory (Bruner, 1966) and with neuroscience-informed accounts of how multiple representational modes support deeper learning (Goldin & Shteingold, 2001; Howard-Jones, 2014).

At a practical level, the findings carry implications for several audiences. For science teachers in madrasah settings, the result indicates that AI tools can be incorporated into regular lessons within the existing timetable provided that students have access to a mobile device and that the tool is structured around the curriculum content rather than offered as an enrichment. For teacher educators, the result suggests that pre-service and in-service training in AI-assisted instruction needs to focus on the pedagogical embedding of AI components rather than on the technical operation of any specific tool (Akgun & Greenhow, 2021; Wasehudin et al., 2024). For policy makers, the finding that an AI-assisted STEM tool can move scientific literacy by approximately 27 raw score points in a four-week trial is sufficient to justify wider piloting, although the policy investment should include teacher training, device access, and content curation rather than the AI tool alone (Wahid, 2024; Zuhriyeh et al., 2025). The historical lineage of mathematics learning reform in Indonesia, from the introduction of realistic mathematics education in the early 2000s (Sembiring et

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

al., 2008) to the contemporary embrace of digital and AI-augmented pedagogies, suggests that durable adoption of innovative pedagogies depends as much on teacher community and institutional support as on the design quality of the tool itself (Azzahra, Diana, & Nuraeni, 2024; Suharja et al., 2024).

Several limitations of the present study should be acknowledged when interpreting the findings. The sample was drawn from a single madrasah, which limits generalisability to other school types and regions. The trial duration of four weeks was sufficient to observe a substantial pretest-to-posttest gain but was too short to assess durability of learning. The outcome measures were researcher-developed instruments calibrated to the trial content; while psychometric properties were acceptable, the instruments may be more sensitive to the specific competencies targeted by the intervention than a standardised test would be. The AI components were operated through a fine-tuned large language model whose responses depend on the underlying training data; the quality of these responses may vary in deployment contexts with different student vocabulary, dialect, or content focus. Finally, the lessons in the experimental group were facilitated by a teacher who had received specific training in the use of STEMI-AI, which may not reflect the implementation fidelity that would be achieved by teachers with only general professional development. These limitations frame the recommendations stated in the conclusion.

CONCLUSION

This study set out to develop, validate, and test the effectiveness of an AI-assisted STEM learning tool (STEMI-AI) for seventh-grade students at a madrasah tsanawiyah in Kudus Regency, Central Java. The three research questions stated in the introduction are answered explicitly below.

In response to the first research question, namely what is the validity of the STEMI-AI learning tool, the result was a mean V Aiken coefficient of 0.91 across four expert validators and six validity dimensions, with the lowest dimension scoring 0.89 and the highest dimension scoring 0.93. All six dimensions exceeded the 0.80 threshold for high validity, and the highest score was observed on pedagogical alignment, which is consistent with the design intent of the tool. The validators provided 23 specific revision suggestions during the validation cycle, of which 19 were incorporated into the third and final iteration of the tool before the trial implementation. The STEMI-AI tool is therefore classified as highly valid in content, construct, language, media presentation, AI scaffold quality, and pedagogical alignment.

In response to the second research question, namely what is the practicality of the STEMI-AI tool, the result was a teacher practicality of 88 percent and a student practicality of 84 percent across seven dimensions of practical use in a four-week trial. The teacher score sits in the highly practical category and the student score sits in the upper end of the practical category. The lowest dimension on the student questionnaire concerned the quality of the adaptive problem generator (82 percent), which points to a specific component for refinement in future iterations. The STEMI-AI tool is therefore classified as practical for use in a real madrasah tsanawiyah classroom.

In response to the third research question, namely what is the effectiveness of the STEMI-AI tool in improving scientific literacy and problem-solving, the result was a statistically significant between-group difference favouring the experimental class on both outcomes. The experimental class reached the medium Hake N-gain range on both outcomes (N-gain = 0.58 for scientific literacy, N-gain = 0.54 for problem-solving), while the control class remained in the low range (N-gain = 0.22 and N-gain = 0.23). The analysis of covariance returned adjusted mean differences of 17.4 points on scientific literacy ($F(1, 61) = 82.34$, $p < .001$) and 15.3 points on problem-solving ($F(1, 61) = 74.18$, $p < .001$), with Cohen's d values of 1.12 and 1.05 respectively. The STEMI-AI tool is therefore classified as effective in improving both target outcomes within a four-week trial implementation. The combined answers to the three research questions support the substantive conclusion that an AI-assisted STEM design is feasible and effective for seventh-grade madrasah tsanawiyah classrooms, provided that the AI components function as pedagogical scaffolds rather than as substitutes for teacher facilitation and provided that the tool is structured around the curriculum content rather than offered as a free-standing enrichment.

Three recommendations follow for future research and practice. First, replications in other madrasah and regional contexts are needed to test the generalisability of the present findings, with particular attention to settings with different device infrastructure and teacher preparedness profiles. Second, longer trial durations of at least one semester are needed to test the durability of learning beyond the immediate posttest, ideally with delayed posttest measurements at four-week, eight-week, and end-of-semester intervals. Third, future iterations of STEMI-AI should focus on the refinement of the adaptive problem generator, which received the lowest practicality score in the present trial and represents the component most likely to benefit from additional fine-tuning with classroom-level student response data. Replication studies might also disaggregate the contribution of the STEM-integration component from the AI-scaffolding component to clarify which design features carry the most weight in the observed gains.

The substantive contribution of this study to the wider literature lies in the demonstration that two research streams that have developed largely in parallel, namely AI in education and STEM integration, can be productively combined within a single design that is implementable in a regular madrasah classroom. The methodological contribution lies in the application of the Plomp three-phase development research model to an AI-assisted educational

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

tool with explicit reporting of validity, practicality, and effectiveness, which provides a transparent template for future Indonesian developers working in similar territory.

Acknowledgments

The authors gratefully acknowledge the headmaster, science teacher, and students of the participating madrasah tsanawiyah in Kudus Regency, Central Java, for their cooperation and dedication throughout the four-week trial implementation. The authors also thank the four expert validators who reviewed the STEMI-AI tool during the prototyping phase and provided detailed revision suggestions. Faiq Makhdum Noor acknowledges the support of Institut Agama Islam Negeri Kudus for the research leave during which this study was conducted. Monera A. Salic-Hairulla acknowledges the support of the Mindanao State University – Iligan Institute of Technology for the collaborative supervision arrangement that made this cross-institutional study possible. Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Conflict of Interest: The authors declare no conflict of interest with respect to the research, authorship, or publication of this article.

REFERENCES

Akgun, S., & Greenhow, C. (2021). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. AI and Ethics, 2(3), 431–440. https://doi.org/10.1007/s43681-021-00096-7

Akker, J. v. d., Bannan, B., Kelly, A. E., Nieveen, N., & Plomp, T. (Eds.). (2013). Educational design research: Part A: An introduction. SLO Netherlands Institute for Curriculum Development.

Antonopoulou, H., Halkiopoulos, C., & Gkintoni, E. (2023). Educational neuroscience and its contribution to math learning. Technium Education and Humanities, 4. https://doi.org/10.47577/teh.v4i.8237

Azzahra, W., Diana, S., & Nuraeni, E. (2024). Unraveling the evolution of brain-based learning in Indonesia: An in-depth exploration through systematic literature review. International Journal of Educational Reform, 33(4), 483–502. https://doi.org/10.1177/10567879241258134

Bang, H. J., Li, L., & Flynn, K. (2023). Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. Early Childhood Education Journal, 51(4), 717–732. https://doi.org/10.1007/s10643-022-01332-3

Bruner, J. S. (1966). Toward a theory of instruction. Harvard University Press.

Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. Computers and Education: Artificial Intelligence, 6, 100197. https://doi.org/10.1016/j.caeai.2023.100197

Cichy, I., Kaczmarczyk, M., Wawrzyniak, S., Kruszwicka, A., Przybyla, T., Klichowski, M., & Rokita, A. (2020). Participating in physical classes using Eduball stimulates acquisition of mathematical knowledge and skills by primary school students. Frontiers in Psychology, 11, 2194. https://doi.org/10.3389/fpsyg.2020.02194

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Coppi, M., Fialho, I., & Cid, M. (2023). Developing a scientific literacy assessment instrument for Portuguese 3rd cycle students. Education Sciences, 13(9), 941. https://doi.org/10.3390/educsci13090941

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. Computers & Education, 210, 104967. https://doi.org/10.1016/j.compedu.2023.104967

Debets, T., Banihashem, S. K., Joosten-Ten Brinke, D., Vos, T. E., de Buy Wenniger, G. M., & Camp, G. (2025). Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. Computers & Education, 234, 105323. https://doi.org/10.1016/j.compedu.2025.105323

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 227, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Dwiputra, D. F. K., Azzahra, W., & Heryanto, F. N. (2023). A systematic literature review on enhancing the success of independent curriculum through brain-based learning innovation implementation. Indonesian Journal on Learning and Advanced Education, 5(3), 262–276. https://doi.org/10.23917/ijolae.v5i3.22318

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191. https://doi.org/10.3758/BF03193146

Goldin, G. A., & Shteingold, N. (2001). Systems of representations and the development of mathematical concepts. In A. A. Cuoco & F. R. Curcio (Eds.), The roles of representation in school mathematics (pp. 1–23). National Council of Teachers of Mathematics.

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

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

Hartoyo, S. R., Nahdi, D. S., & Cahyaningsih, U. (2025). Pengaruh pendekatan realistic mathematics education (RME) berbantuan media ARCA terhadap pemahaman konsep siswa sekolah dasar. STRATEGY: Jurnal Inovasi Strategi dan Model Pembelajaran, 5(3), 348–356. https://doi.org/10.51878/strategi.v5i3.6843

Hidayat, R., & Wardat, Y. (2023). A systematic review of augmented reality in science, technology, engineering and mathematics education. Education and Information Technologies, 28, 11521–11556. https://doi.org/10.1007/s10639-023-12157-x

Howard-Jones, P. A. (2014). Neuroscience and education: Myths and messages. Nature Reviews Neuroscience, 15(12), 817–824. https://doi.org/10.1038/nrn3817

Issholikhah, L. N., Oktradiksa, A., & Shalikhah, N. D. (2024). Realistic mathematics education (RME) model to improve mathematics learning outcomes for MI Muhammadiyah Sriwedari students, Magelang Regency. In Proceedings of the 5th Borobudur International Symposium on Humanities and Social Science 2023 (pp. 793–801). Atlantis Press. https://doi.org/10.2991/978-2-38476-273-6_83

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2022). Kurikulum Merdeka [Independent curriculum]. Government of Indonesia.

Krause, E., Iordan, A. D., Polizzi, A., & Patriotis, D. (2025). The impact of Artificial Intelligence (AI) on students' academic development. Education Sciences, 15(3), 343. https://doi.org/10.3390/educsci15030343

Kurniawan, E. S., Mundilarto, M., & Istiyono, E. (2024). Improving student higher order thinking skills using Synectic-HOTS-oriented learning model. International Journal of Evaluation and Research in Education, 13(2), 1132–1140. https://doi.org/10.11591/ijere.v13i2.25002

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

Liu, D., Tan, X., Yan, H., & Li, W. (2024). Improving mental arithmetic ability of primary school students with schema teaching method: An experimental study. PLOS ONE, 19(4), e0297013. https://doi.org/10.1371/journal.pone.0297013

Liu, M., Ren, Y., Nyagoga, L. M., Stonier, F., Wu, Z., & Yu, L. (2025). The impact of ChatGPT on students' academic achievement: A meta-analysis. Journal of Computer Assisted Learning, 41, e70096. https://doi.org/10.1111/jcal.70096

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282.

Mukuka, A., & Alex, J. K. (2024). Review of research on microteaching in mathematics teacher education: Promises and challenges. Eurasia Journal of Mathematics, Science and Technology Education, 20(1), em2381. https://doi.org/10.29333/ejmste/13941

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. British Journal of Educational Technology, 55(4), 1328–1353. https://doi.org/10.1111/bjet.13454

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. Educational Studies in Mathematics, 102(1), 9–28. https://doi.org/10.1007/s10649-019-09903-9

Nur, S., & Ikhsan, J. (2024). Implementation of STEM integrated problem based learning model to improve problem solving skills and learning motivation of grade X vocational high school students on the material of substances and their changes. Jurnal Penelitian Pendidikan IPA, 10(11), 8882–8891. https://doi.org/10.29303/jppipa.v10i11.9121

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

Plomp, T., & Nieveen, N. (Eds.). (2013). Educational design research. SLO Netherlands Institute for Curriculum Development.

Polya, G. (1957). How to solve it: A new aspect of mathematical method (2nd ed.). Princeton University Press.

Risdiyanti, I., Zulkardi, Z., Putri, R. I. I., Prahmana, R. C. I., & Nusantara, D. S. (2024). Ratio and proportion through realistic mathematics education and pendidikan matematika realistik Indonesia approach: A systematic literature review. Jurnal Elemen, 10(1). https://doi.org/10.29408/jel.v10i1.24445

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohana, S., Irianto, A., & Rachmadtullah, R. (2023). Project-based learning model on critical thinking ability seen from cognitive style in elementary schools. Studies in Learning and Teaching, 4(2), 245–259.

Rojas, M., Nussbaum, M., Chiuminatto, P., Guerrero, O., Greiff, S., Krieger, F., & Van Der Westhuizen, L. (2021). Assessing collaborative problem-solving skills among elementary school students. Computers & Education, 175, 104313. https://doi.org/10.1016/j.compedu.2021.104313

e-ISSN: 2987-2421
https://doi.org/10.37630/bijee.v4i1.4447

Säfström, A. I., Lithner, J., Palm, T., Palmberg, B., Sidenvall, J., Andersson, C., Boström, E., & Granberg, C. (2024). Developing a diagnostic framework for primary and secondary students' reasoning difficulties during mathematical problem solving. Educational Studies in Mathematics, 115(1), 125–149. https://doi.org/10.1007/s10649-023-10278-1

Samritin, S., Natsir, S. R., Manaf, A., & Sari, E. R. (2023). The effect of realistic mathematics education implementation in mathematics learning in elementary school. Formatif: Jurnal Ilmiah Pendidikan MIPA, 13(1), 81–88. https://doi.org/10.30998/formatif.v13i1.16522

Schoenfeld, A. H. (1985). Mathematical problem solving. Academic Press.

Sembiring, R. K., Hadi, S., & Dolk, M. (2008). Reforming mathematics learning in Indonesian classrooms through RME. ZDM, 40, 927–939. https://doi.org/10.1007/s11858-008-0125-9

Suanto, E., Maat, S. M., & Zakaria, E. (2023). The effectiveness of the implementation of three dimensions geometry KARA module on higher order thinking skills (HOTS) and motivation. International Journal of Instruction, 16(3), 95–116. https://doi.org/10.29333/iji.2023.1636a

Suharja, S., Mustadi, A., & Oktari, V. (2024). Examining brain based learning models assisted open-ended approach to mathematics understanding concept. Jurnal Prima Edukasia, 12(1), 19–29. https://doi.org/10.21831/jpe.v12i1.67303

Susanti, E. (2025). Enhancing problem-solving skills in elementary students through realistic mathematics education. SCIENCE: Jurnal Inovasi Pendidikan Matematika dan IPA, 5(1), 48. https://doi.org/10.51878/science.v5i1.4344

Vuong, Q. A., Bui, D. T., Dang, H. T. T., Le, A. V., & Do, D. L. (2025). Teachers' perspectives on the application of technology in mathematics education in primary schools: A dataset from Vietnam. Data in Brief, 60, 111473. https://doi.org/10.1016/j.dib.2025.111473

Wahdaniyah, N., Agustini, R., & Tukiran, T. (2023). Analysis of effectiveness PBL-STEM to improve student's critical thinking skills. IJORER: International Journal of Recent Educational Research, 4(3), 365–382. https://doi.org/10.46245/ijorer.v4i3.312

Wahid, S. H. (2024). Exploring the intersection of Islam and digital technology: A bibliometric analysis. Social Sciences & Humanities Open, 10, 101085. https://doi.org/10.1016/j.ssaho.2024.101085

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. Humanities and Social Sciences Communications, 12, 621. https://doi.org/10.1057/s41599-025-04787-y

Wasehudin, W., Hidayat, T., Hadijah, S., Hairullah, H., & Kosim, A. (2024). Artificial Intelligence's impact on the development of Islamic Religious Education learning at a public junior high school of Cilegon, Indonesia. Hanifiya: Jurnal Studi Agama-Agama, 7(2), 193–198.

Wen, T., He, J., & Yang, Y. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. Disciplinary and Interdisciplinary Science Education Research, 5, 21. https://doi.org/10.1186/s43031-023-00093-2

Zhao, X., Feng, G. C., Ao, S. H., & Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. BMC Medical Research Methodology, 22, 232. https://doi.org/10.1186/s12874-022-01707-5

Zuhriyeh, S., Ali, M., & Hidayat, A. (2025). Digital transformation of Islamic education: An artificial intelligence-based teaching module development study. Edunesia: Jurnal Ilmiah Pendidikan, 6(2), 1113–1126. https://doi.org/10.51276/edu.v6i2.1255