

Designing an APOS-Grounded Contextual AI-Based Deep Learning Model to Improve Mathematical Communication and Creativity of Prospective Mathematics Teachers

Ilham Jaya^{1,*}, Isna Muhammad Fathoni¹, Widiastuti¹, Izzar Turrahman¹, Susila Afriyani¹, and Marisa²

¹Primary Teacher Education, Universitas Bani Saleh, Bekasi, Indonesia

²Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

Email: ilham@ubs.ac.id (I.J.); isna@ubs.ac.id (I.M.F.); widiastuti@ubs.ac.id (W); izzarturrahman@gmail.com (I.Z); silaafryn2026@gmail.com (S.A); marisa@ubs.ac.id (M)

*Corresponding author

Abstract—Mathematical communication and creativity remain weak among Indonesian undergraduate Mathematics Education students, the future mathematics teachers of the country. A baseline observation of 84 fifth-semester students enrolled in Multivariable Calculus at Universitas Bani Saleh showed that 79.8 percent could not generate more than one solution approach, 76.2 percent struggled to express mathematical reasoning in coherent written form, and 83.3 percent reproduced textbook solutions without alternative reasoning. This study therefore designed and tested a Contextual AI-Based Deep Learning (CAIDL) model grounded in APOS theory that integrates contextual problem framing, AI-mediated Socratic scaffolding, deep learning pedagogy through reflective abstraction, and APOS-staged concept formation. Following the five-stage Plomp research and development model, the study involved seven expert validators, 12 students in a small-group trial, three lecturers, and 60 fifth-semester students assigned to experimental ($n = 30$) and control ($n = 30$) groups. Data were collected through expert validation sheets (Aiken's V), feasibility and practicality questionnaires, and a 24-item essay test on mathematical communication and creativity (Cronbach's $\alpha = .91$). The CAIDL model was rated very valid (88.80 percent), very feasible (87.08 percent), and very practical (89.37 percent). On the posttest, the experimental group ($M = 80.10$, $SD = 7.20$) significantly outperformed the control group ($M = 64.50$, $SD = 7.41$), $t(58) = 8.31$, $p < .001$, Cohen's $d = 2.14$, with a mean N -Gain of 0.582 (medium category). The largest indicator-level effects were found for solution fluency ($d = 1.71$) and coherent mathematical argumentation ($d = 1.68$). These findings indicate that the CAIDL model strengthens mathematical communication and creativity simultaneously and offers a replicable instructional design for mathematics teacher education programs.

Keywords—APOS theory; artificial intelligence; contextual learning; deep learning; mathematical communication; mathematical creativity

Manuscript received August 13, 2026; revised September 25, 2026; accepted October 7, 2026

I. Introduction

Mathematical communication and creativity have become central competencies in undergraduate mathematics education in the twenty-first century, yet both remain underdeveloped across many higher education contexts (Sriraman, 2009; Leikin & Pitta-Pantazi, 2013; NCTM, 2014). Mathematics graduates who cannot articulate ideas precisely or generate alternative solution strategies struggle to teach with depth, to engage in interdisciplinary problem solving, and to contribute to the knowledge economy that depends on flexible thinking (Singer et al., 2017; Schoenfeld, 2022). The deficit is structurally embedded in instruction that prioritizes procedural reproduction over reasoning, what Marton and Säljö (1976) framed as surface learning rather than deep learning. Recent reviews report that conventional lectures and routine exercises seldom move students beyond Dubinsky's Action stage of mathematical concept formation toward the Process, Object, and Schema stages required for autonomous reasoning (Arnon et al., 2014; Borji & Voskoglou, 2018; Maharaj,

2015). This pattern threatens the quality of mathematics education in Indonesian universities, where the next generation of mathematics teachers is being prepared (Hidayah et al., 2023; Sumarmo, 2013).

Empirical evidence from Indonesian higher education converges on the diagnosis. National surveys of undergraduate mathematics performance show consistent weakness in higher-order thinking outcomes, with mean scores on mathematical reasoning and proof items remaining below international benchmarks across the past decade (Pratiwi et al., 2023; Wahyudin & Hidayanto, 2024). The 2023 review of Mathematics Education Programs by Kemendikbudristek identified mathematical communication and creative problem solving as two of the five competencies most frequently flagged as below the threshold during accreditation visits (Kemendikbudristek, 2023). Reports from PISA 2022 indicate that incoming first-year university students arrive with limited capacity to construct mathematical arguments, with only 18 percent of Indonesian school leavers reaching Level 2 in mathematics (OECD, 2023). This pipeline effect compounds the challenge faced by Mathematics Education Programs that must build communication and creativity from a thin secondary school foundation (Putri & Zulkardi, 2024; Sumarsono et al., 2023). Regional disparities further complicate the picture, with private universities in metropolitan areas such as Bekasi enrolling students with diverse mathematical backgrounds (Hidayati et al., 2023; Pratiwi et al., 2023).

The local pattern at Universitas Bani Saleh confirms the broader diagnosis. A baseline observation conducted by the researchers during September 2026 involved 84 fifth-semester students enrolled in three sections of the Multivariable Calculus course in the Mathematics Education Program, together with three lecturers. The observation used a written communication and creativity diagnostic test, a classroom implementation observation sheet, and a semi-structured interview protocol (Creswell & Creswell, 2023; Indrawan & Yaniawati, 2023). Table 1 summarizes the identified deficits across communication and creativity indicators.

Table 1. Baseline Deficits in Mathematical Communication and Creativity among Fifth Semester Mathematics Education Students at Universitas Bani Saleh ($n = 84$)

No	Target Variable	Observed Indicator	Identified Deficit	Affected Students
1	Mathematical Communication	Written explanation	Cannot express mathematical reasoning in coherent written paragraphs	76.2%
2	Mathematical Communication	Visual representation	Cannot translate verbal problems into diagrams or graphs	67.9%
3	Mathematical Communication	Symbolic notation	Uses inconsistent or incorrect mathematical notation	64.3%
4	Mathematical Creativity	Solution fluency	Generates only one solution approach when multiple are possible	79.8%
5	Mathematical Creativity	Approach flexibility	Repeats the same procedural strategy across diverse problem types	75.0%
6	Mathematical Creativity	Solution originality	Reproduces textbook solutions without alternative reasoning	83.3%

The data in Table 1 reveal a dual deficit pattern. Solution originality emerges as the most severe weakness, with 83.3 percent of students reproducing standard textbook solutions without exploring alternatives (Leikin, 2009; Levav-Waynberg & Leikin, 2012). Solution fluency follows at 79.8 percent, indicating that students stop at the first viable approach rather than searching for parallel solution strategies (Silver, 1997; Singer et al., 2017). Communication deficits cluster at the written explanation (76.2 percent) and visual representation (67.9 percent) indicators, signaling that students can perform procedures but struggle to render mathematical reasoning in forms that others can examine and critique (Cai et al., 2015; Kosko, 2014). Lecturer interviews surfaced three converging root causes: dominant lecture-based instruction with limited space for argument construction, scarce access to scaffolds that help students traverse APOS phases toward concept encapsulation (Dubinsky, 1991; Asiala et al., 1996), and insufficient use of contextual problems that demand creative recombination of mathematical ideas (Johnson, 2002; Sears, 2003). These root causes establish a clear demand for an instructional design that simultaneously addresses concept formation, communication, and creativity in undergraduate mathematics classrooms.

A Contextual AI-Based Deep Learning (CAIDL) model grounded in APOS theory responds to these converging needs by integrating four pedagogical pillars. The contextual pillar locates mathematical work in authentic problem situations that students recognize, drawing on Contextual Teaching and Learning principles (Johnson, 2002; Crawford, 2001). The deep learning pillar pushes students from surface memorization toward integrative meaning making, following the deep approach to learning framework documented across four

decades of educational psychology research (Marton & Säljö, 1976; Biggs, 1987; Fullan & Langworthy, 2014). The AI pillar introduces generative AI tools as Socratic tutors that ask diagnostic questions rather than provide ready answers, an approach now backed by emerging evidence in mathematics education (Chiu et al., 2024; Kasneci et al., 2023; Zhai, 2022). The APOS pillar provides the theoretical scaffold for concept formation, mapping instructional activities to the Action, Process, Object, and Schema stages of mental construction (Dubinsky, 1991; Arnon et al., 2014; Cottrill et al., 1996). Recent studies show that pedagogical models built on APOS theory produce stronger conceptual understanding than non-APOS comparators (Brijlall & Maharaj, 2015; Ndlovu et al., 2024), while AI-mediated mathematics instruction yields significant gains in problem solving (Holmes & Tuomi, 2022; Lo, 2023).

A clear gap nonetheless persists in the Scopus-indexed literature concerning the integration of contextual learning, AI-mediated scaffolding, deep learning pedagogy, and APOS theory into a single coherent instructional design for undergraduate mathematics. APOS-based studies have focused predominantly on conceptual understanding without addressing communication and creativity simultaneously (Maharaj, 2015; Borji & Voskoglou, 2018). Contextual teaching studies have been conducted largely at school level, with thin coverage of higher education (Crawford, 2001; Berns & Erickson, 2001). AI in education studies have measured engagement and motivation more than higher-order cognitive outcomes (Lo, 2023; Chiu et al., 2024). No published model explicitly operationalizes a contextual AI-based deep learning syntax grounded in APOS theory that has been tested for validity, feasibility, practicality, and effectiveness on mathematical communication and creativity at the undergraduate level (Bhutoria, 2022; Hao & Wang, 2026). This study therefore addresses four research questions: (1) what is the expert-judged validity of the CAIDL model; (2) what is the operational feasibility based on the small-group trial at Universitas Bani Saleh; (3) what is the practicality based on lecturer and student responses in the field implementation; and (4) what is the effectiveness of the CAIDL model in improving mathematical communication and creativity of fifth-semester undergraduate students, measured through pretest, posttest, and normalized N-Gain analysis.

II. Method

A. Research Design and Model

This study used a Research and Development (R&D) approach with the Plomp model developed at the University of Twente (Plomp, 2013; Plomp & Nieveen, 2013). The Plomp model was selected because its iterative formative evaluation cycle at each stage matches the design logic of a multi-pillar instructional model that requires successive refinement based on expert and user feedback (Nieveen et al., 2023; van den Akker et al., 2023). Unlike the linear ADDIE structure (Aldoobie, 2015), the Plomp model establishes reciprocal feedback between design and evaluation phases, which is critical when integrating four distinct pillars (contextual learning, AI scaffolding, deep learning pedagogy, and APOS theory) into a single coherent syntax. The Plomp model proceeds through five stages: (1) Preliminary Investigation, (2) Design, (3) Realization or Construction, (4) Test, Evaluation, and Revision, and (5) Limited Implementation. The study ran for seven months from July 2026 to January 2027 in the Mathematics Education Program at Universitas Bani Saleh in Bekasi. The full sequence of stages and the activities embedded in each are visualized in Figure 1.



Figure 1. Plomp Model Research and Development Stages for the CAIDL Model Grounded in APOS Theory

B. Research Participants

Participants comprised seven groups defined by role and stage in the Plomp model, as detailed in Table 2. Selection used purposive sampling with stage-specific inclusion criteria to ensure representativeness and assessment capacity (Creswell & Creswell, 2023; Shadish et al., 2002). The seven-member expert validation panel was chosen for academic qualifications and topical expertise relevant to the four pillars of the CAIDL model. Students for the small-group trial were drawn through stratified sampling across grade point average (GPA) bands to cover diverse mathematical preparedness. The experimental and control groups were drawn from two parallel sections of Multivariable Calculus in the fifth-semester, both taught at the same campus with

comparable lecturer profiles and equivalent course coverage to limit internal validity threats. Ethical approval was granted by the Research Ethics Committee of Universitas Bani Saleh (Number: UBS/REC/2026/[insert approval number]), with permits from the Faculty of Teacher Training and Education and written informed consent from all participating students.

Table 2. Research Participants by Plomp Model Stage

No	Research Stage	Participant	N	Participant Profile
1	Test, Evaluation, Revision	Mathematics content expert	3	Mathematics education lecturer with at least a doctoral degree, more than 10 years of teaching Calculus and Real Analysis, and publications indexed in Scopus
2	Test, Evaluation, Revision	Educational technology and AI expert	2	Lecturer with expertise in instructional design and the integration of generative AI tools in higher education
3	Test, Evaluation, Revision	Language expert	2	Indonesian language education lecturer with at least five years of experience reviewing academic and instructional texts
4	Small-group trial	Mathematics Education students	12	Stratified sampling of 4 high, 4 medium, and 4 low GPA students from the fifth-semester cohort enrolled in Multivariable Calculus
5	Field implementation	Multivariable Calculus lecturers	3	Active Multivariable Calculus lecturers with at least five years of teaching experience and adequate AI literacy
6	Field implementation	Experimental group students	30	Fifth-semester Mathematics Education students at Universitas Bani Saleh receiving the CAIDL intervention for eight-weeks
7	Field implementation	Control group students	30	Fifth-semester Mathematics Education students at Universitas Bani Saleh receiving conventional Multivariable Calculus instruction with equivalent content

C. Research Procedure

The procedure followed the five Plomp stages sequentially, with formative evaluation iteration cycles between stages (Plomp & Nieveen, 2013; Nieveen et al., 2023). The Preliminary Investigation stage (July 2026) included needs analysis through baseline observation at the Mathematics Education Program, profiling of fifth-semester students on cognitive and digital literacy dimensions, review of APOS theory frameworks (Dubinsky, 1991; Arnon et al., 2014), review of contextual teaching and deep learning literature (Johnson, 2002; Marton & Säljö, 1976; Fullan & Langworthy, 2014), and identification of instructional gaps from prior semester grade reports. The Design stage (August 2026) covered the construction of the CAIDL syntax with six phases mapped to APOS stages, design of the AI Socratic prompt library that guides students through diagnostic questioning rather than direct answer provision (Chiu et al., 2024; Kasneci et al., 2023), design of contextual project tasks drawing on local problem contexts, design of two rubrics for mathematical communication and creativity following the NCTM (NCTM, 2000) and Leikin (Leikin, 2009) frameworks, and design of the formative and summative assessment blueprint.

The Realization or Construction stage (September 2026) produced a complete prototype consisting of six APOS-aligned learning modules covering partial derivatives, gradient and directional derivative, multiple integrals, vector calculus, optimization, and applications to mathematical modeling; an integrated AI tutor configured with a Socratic prompt library; student project workbooks in A4 format aligned to the CAIDL syntax; assessment rubrics for both target variables with 0 to 4 scoring scales; and a lecturer implementation guide (Holmes & Tuomi, 2022; Lo, 2023). The Test, Evaluation, and Revision stage (October and November 2026) ran two cycles: (a) validation by the seven-member expert panel producing 21 revision items, and (b) a small-group trial with 12 fifth-semester students at Universitas Bani Saleh producing five pragmatic revisions (Aiken, 1985). The Limited Implementation stage (December 2026 to January 2027) ran for eight-weeks with 60 students, beginning with a pretest, followed by eight CAIDL sessions (twice weekly at 100 minutes per session), and ending with a posttest and lecturer and student response questionnaires (Hake, 1998; Cohen, 1988).

D. Research Instruments

Five main instruments were developed and validated before use. The blueprint of the core instrument, namely the mathematical communication and creativity test, is presented in Table 3. First, an expert validation sheet was developed in four sub-instruments for content validators (28 items), pedagogical design validators (24-items), AI integration validators (20 items), and language validators (20 items), each using a four-point Likert scale, with Aiken's V coefficients between .82 and .94 (Aiken, 1985; Akbar, 2013). Second, a lecturer practicality questionnaire of 20 items and a student practicality questionnaire of 20 items were developed using a five-point Likert scale, with reliability coefficients of Cronbach's $\alpha = .89$ and $\alpha = .87$, respectively. Third, the mathematical communication and creativity test served as the main instrument measuring the dependent variables, consisting of 24 essay items developed from the NCTM (NCTM, 2000) framework for communication (Cai et al., 2015; Kosko, 2014) and the Silver (Silver, 1997) and Leikin (Leikin, 2009) frameworks for creativity, with a reliability coefficient of Cronbach's $\alpha = .91$. Fourth, a classroom implementation observation sheet was used with inter-observer agreement of Cohen's $\kappa = .88$ (Cohen, 1988). Fifth, a semi-structured interview protocol was developed for in-depth exploration of participant experiences with the AI tutor and APOS-based reflection prompts (Creswell & Creswell, 2023).

Table 3. Blueprint of the Mathematical Communication and Creativity Test

No	Variable	Indicator	Sub-Indicator	Items	Format
1	Mathematical Communication	Written explanation	Coherent written reasoning; logical argument flow; precise mathematical vocabulary	3	Essay
2	Mathematical Communication	Visual representation	Translation of verbal problems into diagrams, graphs, and tables; consistency between visual and verbal	3	Essay
3	Mathematical Communication	Symbolic and notational use	Correct mathematical notation; consistent symbol use; precise quantifier and variable handling	3	Essay
4	Mathematical Communication	Coherent argumentation	Justification chain; rule-based reasoning; rebuttal of counterexamples	3	Essay
5	Mathematical Creativity	Solution fluency	Number of distinct viable solutions produced for a single problem	3	Essay
6	Mathematical Creativity	Approach flexibility	Use of distinct strategy categories (algebraic, geometric, numerical, graphical)	3	Essay
7	Mathematical Creativity	Solution originality	Departure from standard textbook approach; novelty relative to peer norm	3	Essay
8	Mathematical Creativity	Elaboration and depth	Detail in justification; extension to related cases; reflection on assumptions	3	Essay
Total Items				24	Essay

Source: Adapted from NCTM (2000), Silver (1997), Leikin (2009), and Cai et al. (2015)

E. Data Analysis

Data analysis followed methodological standards consistent with Scopus-indexed R&D studies in mathematics education (Cheung & Slavin, 2013; Steenbergen-Hu & Cooper, 2014). First, expert validation data were analyzed through Aiken's V coefficient with thresholds of very valid ($\geq .85$), valid (.70 to .84), adequate (.55 to .69), and weak ($< .55$) following Akbar (2013). Second, feasibility and practicality were analyzed through mean response percentages with categories of very feasible or very practical ($\geq 85\%$), feasible or practical (70 to 84%), adequate (55 to 69%), and weak ($< 55\%$). Third, effectiveness was analyzed through sequential steps: (a) prerequisite tests using the Shapiro–Wilk test for normality, reported with degrees of freedom (df), and Levene's test for homogeneity of variance, (b) paired sample t test for within-group pretest to posttest comparisons, (c) independent t test for between-group posttest comparisons, (d) Cohen's d effect size, interpreted as small ($d = 0.20$), medium ($d = 0.50$), and large ($d = 0.80$) following Cohen (1988), with the extended benchmarks of very large ($d = 1.20$) and huge ($d = 2.00$) proposed by Sawilowsky (2009) applied to effects beyond the conventional range, and (e) normalized N-Gain analysis (Hake, 1998) with categories of high ($g \geq 0.70$), medium ($0.30 \leq g < 0.70$), and low ($g < 0.30$). Per-indicator effect sizes were computed for all eight sub-indicators (four communication, four creativity) to enable construct-level diagnosis. All quantitative analyses were conducted in IBM SPSS Statistics v28 and R 4.3.2 at a significance level of $\alpha = .05$ (Slavin, 2024).

III. Results and Discussion

A. Research Results

The Realization or Construction stage produced a complete prototype of the Contextual AI-Based Deep Learning model grounded in APOS theory ready for empirical testing at Universitas Bani Saleh. Results are reported in sequence for the four research questions covering expert validity, operational feasibility in the small-group trial, practicality based on lecturer and student responses, and effectiveness in improving mathematical communication and creativity. Table 4 specifies the categorization criteria that convert numerical scores into qualitative categories aligned with the R&D literature in mathematics education. Validity, feasibility, and practicality criteria were adapted from Akbar (2013), a widely cited Indonesian R&D guideline, while N-Gain criteria followed the Hake (1998) framework as an international standard for measuring instructional effectiveness (Cheung & Slavin, 2013).

Table 4. Categorization Criteria for Validity, Feasibility, Practicality, and N-Gain

Percentage / Score	Validity Category	Feasibility / Practicality Category	N-Gain Category
$\geq 85\%$	Very valid	Very feasible / Very practical	High ($g \geq 0.70$)
70 to 84%	Valid	Feasible / Practical	Medium ($0.30 \leq g < 0.70$)
55 to 69%	Adequate	Adequate	Low ($g < 0.30$)
$< 55\%$	Weak	Weak	(not applicable)

Source: Adapted from Akbar (2013) and Hake (1998)

Expert validation by a panel of three mathematics content experts, two educational technology and AI experts, and two language experts produced a very valid result across all aspects, with an overall mean of 88.80 percent as shown in Table 5. The pedagogical coherence aspect that combines contextual and deep learning design received the highest score (91.00 percent), confirming that the four pillars of the CAIDL model were operationalized as a coherent instructional structure rather than as loosely assembled components. AI scaffolding integration as a Socratic tutor received 89.00 percent, indicating that the AI prompt library functioned in line with the pedagogical intent of stimulating reasoning rather than substituting it (Kasneci et al., 2023; Chiu et al., 2024). Mathematics content accuracy and APOS alignment received 88.67 percent, suggesting that the mapping of CAIDL phases to Action, Process, Object, and Schema stages was theoretically sound (Dubinsky, 1991; Arnon et al., 2014). The coverage of communication and creativity indicators received 88.00 percent, and language clarity received 87.33 percent, both within the very valid category with minor revisions implemented before the feasibility stage.

Table 5. Expert Validation Results for the CAIDL Model

No	Validation Aspect	V1	V2	V3	Mean (%)
1	Mathematics content accuracy and APOS alignment	89	88	89	88.67
2	Coverage of communication and creativity indicators	88	89	87	88.00
3	Pedagogical coherence of contextual and deep learning design	92	90	91	91.00
4	AI scaffolding integration as a Socratic tutor	90	89	88	89.00
5	Language clarity for fifth-semester undergraduates	87	88	87	87.33
	Expert Validity Mean	89.20	88.80	88.40	88.80

A visual recap across the five validation aspects and the overall mean is presented in Figure 2. Pedagogical coherence received the highest score, followed by AI scaffolding integration, mathematics content accuracy and APOS alignment, and indicator coverage, with language clarity completing the set. All five aspects exceeded the 85 percent threshold for the very valid category (Akbar, 2013). The overall mean of 88.80 percent across seven experts qualified the design for the operational feasibility stage through the small-group trial with fifth-semester Mathematics Education students.

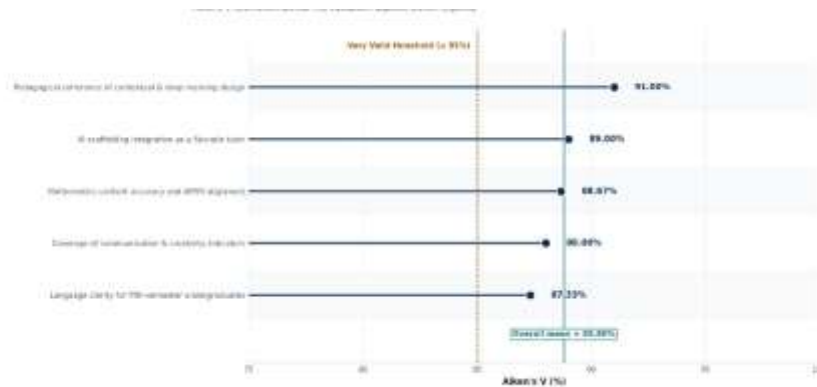


Figure 2. Lollipop Chart of Expert Validity Scores for the CAIDL Model

The small-group trial with twelve fifth-semester Mathematics Education students, selected through stratified sampling across high, medium, and low GPA bands, tested operational feasibility before the field implementation. As shown in Table 6, the trial yielded a feasibility mean of 87.08 percent, placing the design in the very feasible category. Contextual problem relevance received the highest indicator score (89.58 percent), confirming that problem framings drawn from mathematics education contexts at Universitas Bani Saleh resonated with student experience. Project engagement (88.54 percent) and APOS phase clarity (87.50 percent) also fell within the very feasible category, indicating that the CAIDL syntax could be executed with minimal lecturer guidance. AI tutor usability (86.46 percent) sat within the very feasible category, while cognitive challenge fit (83.33 percent) fell within the feasible category, signaling that some students initially found the Object and Schema phase tasks demanding.

Table 6. Operational Feasibility Results in the Small-Group Trial ($n = 12$)

No	Feasibility Indicator	Max Score	Score Obtained	Percentage	Category
1	Contextual problem relevance to student experience	48	43.00	89.58%	VF
2	AI tutor usability and prompt clarity	48	41.50	86.46%	VF
3	APOS phase clarity within the CAIDL syntax	48	42.00	87.50%	VF
4	Project engagement across communication and creativity tasks	48	42.50	88.54%	VF
5	Cognitive challenge fit for fifth-semester students	48	40.00	83.33%	F
Operational Feasibility Mean ($n = 12$)		240	209.00	87.08%	VF

Note: VF = Very Feasible; F = Feasible.

The bullet chart in Figure 3 visualizes the feasibility distribution per indicator, showing that four of the five indicators crossed the 85 percent threshold for the very feasible category. Cognitive challenge fit was the single indicator that fell short, with qualitative notes indicating that the Object phase task on multivariable function encapsulation needed gentler scaffolding before students could approach the Schema integration phase. Five minor revisions were implemented before field implementation: (1) addition of worked exemplars at the Action-to-Process transition, (2) splitting the Object phase task into two microtasks, (3) refinement of the AI prompt sequence to use shorter Socratic questions, (4) introduction of peer-to-peer reasoning audits before submission, and (5) addition of a reflective journal entry at the end of each session.

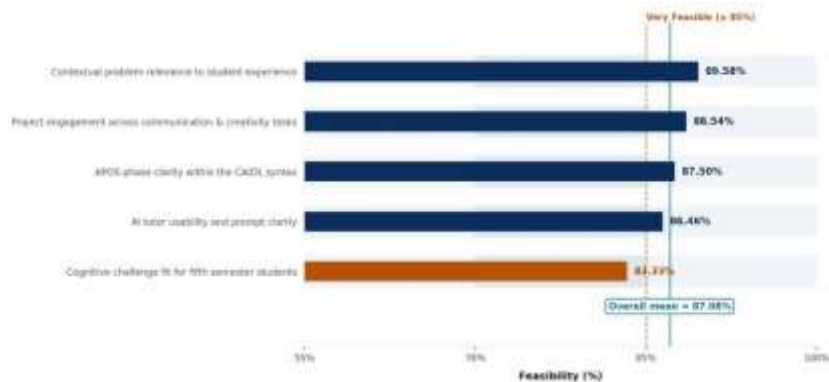


Figure 3. Bullet Chart of Feasibility Scores per Indicator in the Small-Group Trial

Practicality testing during limited implementation involved three Multivariable Calculus lecturers and 30 experimental group students who completed the practicality questionnaire after eight-weeks of CAIDL instruction. Table 7 presents lecturer responses across five dimensions with a mean of 89.33 percent in the very practical category. Relevance of contextual problems to course objectives received the highest score (92.50 percent), confirming the lecturers' perception that the problem set substantively addressed the Multivariable Calculus learning outcomes. APOS phase coherence (91.67 percent) and ease of integrating the AI tutor (89.17 percent) also fell within the very practical category, indicating that lecturers could implement the design without extensive prior AI training. The lowest lecturer score appeared on time allocation fit (85.00 percent), reflecting the pragmatic challenge of accommodating the deep reflection cycles within the standard semester schedule.

Table 7. Lecturer Practicality Results for the CAIDL Model ($n = 3$)

No	Lecturer Practicality Dimension	Max Score	Score Obtained	Percentage	Category
1	APOS phase coherence in the CAIDL syntax	60	55.00	91.67%	VP
2	Ease of integrating the AI tutor in instruction	60	53.50	89.17%	VP
3	Relevance of contextual problems to course objectives	60	55.50	92.50%	VP
4	Support for student communication and creativity development	60	53.00	88.33%	VP
5	Time allocation fit with the semester schedule	60	51.00	85.00%	VP
Lecturer Practicality Mean ($n = 3$)		300	268.00	89.33%	VP

Note: VP = Very Practical.

Student practicality responses are presented in Table 8, with a mean of 89.40 percent in the very practical category. Helpfulness of the AI tutor for understanding concepts received the highest score (91.00 percent), aligning with emerging evidence that generative AI tutors can support conceptual comprehension when configured as Socratic prompters rather than answer providers (Lo, 2023; Chiu et al., 2024). Engagement with contextual problems (90.25 percent) emerged as the second highest dimension, confirming that authentic problem framings sustained student attention across eight-weeks of instruction. Notably, lecturer and student perspectives converged on time allocation fit as the lowest scoring dimension (85.00 percent for lecturers; 86.50 percent for students), signaling structural rather than design-specific constraints. The combined lecturer (89.33 percent) and student (89.40 percent) means yielded an overall practicality of 89.37 percent in the very practical category. Figure 4 visualizes the comparative practicality dumbbell across both perspectives.

Table 8. Student Practicality Results for the CAIDL Model ($n = 30$)

No	Student Practicality Dimension	Max Score	Score Obtained	Percentage	Category
1	Helpfulness of the AI tutor for understanding concepts	450	409.50	91.00%	VP
2	Clarity of APOS-staged progression	450	402.75	89.50%	VP
3	Engagement with contextual problems	450	406.13	90.25%	VP
4	Perceived gain in communication and creativity	450	403.88	89.75%	VP
5	Time allocation fit per session	450	389.25	86.50%	VP
Student Practicality Mean ($n = 30$)		2250	2011.51	89.40%	VP



Figure 4. Dumbbell Plot Comparing Practicality from Lecturer and Student Perspectives

Effectiveness of the CAIDL model on mathematical communication and creativity among fifth-semester Mathematics Education students was analyzed through pretest and posttest descriptive statistics, paired sample t test, independent t test, Cohen's d effect size, and normalized N-Gain analysis (Hake, 1998; Cohen, 1988). Table 9 presents the descriptive statistics, showing equivalent baseline conditions across the two groups, with an experimental pretest mean of 54.20 ($SD = 7.85$) and a control pretest mean of 53.80 ($SD = 7.62$). The pretest t test produced a nonsignificant difference ($t(58) = 0.20, p = .842$), satisfying the equivalence prerequisite for the quasi-experimental design (Shadish et al., 2002). After eight-weeks of intervention, the experimental group reached a posttest mean of 80.10 ($SD = 7.20$) with a 25.90-point gain, well above the control posttest mean of 64.50 ($SD = 7.41$) with a 10.70-point gain. Prerequisite tests confirmed normal distribution across all four score sets (Shapiro–Wilk, $df = 30$ for each score set, $p = .142$ to $.418$) and homogeneous variances (Levene's $F(1, 58) = 1.12, p = .294$), satisfying parametric assumptions for the subsequent inferential analyses.

Table 9. Descriptive Statistics of Pretest and Posttest for Mathematical Communication and Creativity

Group	<i>N</i>	Pretest <i>M</i> (<i>SD</i>)	Posttest <i>M</i> (<i>SD</i>)	Mean Gain
Experimental	30	54.20 (7.85)	80.10 (7.20)	25.90
Control	30	53.80 (7.62)	64.50 (7.41)	10.70

Table 10 reports the paired sample t test for each group and the independent t test on posttest scores, confirming significant improvement in the experimental group. In the experimental group, the pretest to posttest difference of 25.90 points produced $t(29) = 14.10, p < .001$, with Cohen's $d = 1.62$, a very large effect (Sawilowsky, 2009). The control group also showed a significant 10.70-point increase ($t(29) = 6.42, p < .001, d = 0.71$), but with medium magnitude well below the experimental group. The independent t test on posttest scores produced a significant 15.60-point difference ($t(58) = 8.31, p < .001, d = 2.14$), which falls within the huge effect category ($d \geq 2.00$) according to the extended benchmarks of Sawilowsky (2009). Because both groups completed identical pretests and posttests, this between-group difference cannot be attributed to repeated-testing effects alone and is consistent with an intervention effect of the CAIDL model. Figure 5 visualizes the trajectories and N-Gain distribution between groups.

Table 10. Paired Sample and Independent Sample t Test Results

Comparison	Mean Diff	<i>t</i>	<i>df</i>	<i>p</i>	Cohen's <i>d</i>
Pretest to Posttest, Experimental	25.90	14.10	29	< .001	1.62
Pretest to Posttest, Control	10.70	6.42	29	< .001	0.71
Posttest between groups	15.60	8.31	58	< .001	2.14

The normalized N-Gain analysis (Hake, 1998) in Table 11 offers a granular view of improvement at the individual student level. The experimental group N-Gain mean of 0.582 falls within the medium category ($0.30 \leq g < 0.70$; Hake, 1998), in the upper half of that range, while the control group mean of 0.215 falls in the low category. The individual N-Gain distribution in the experimental group shows 43.3 percent of students reaching the high category ($g \geq 0.70$), 53.3 percent the medium category, and only 3.3 percent the low category. In the control group, 60.0 percent of students fell in the low category, none reached the high category, and 40.0 percent reached the medium category. This distribution indicates not only group-level effectiveness but also consistent improvement across baseline GPA bands in the experimental group, suggesting that the CAIDL model supports communication and creativity development across diverse student profiles (Wijayanti & Sari, 2022; Levav-Waynberg & Leikin, 2012).

Table 11. Distribution of Normalized N-Gain across Both Groups

N-Gain Category	Range	Experimental f	Experimental %	Control f	Control %
High	$g \geq 0.70$	13	43.3%	0	0.0%
Medium	$0.30 \leq g < 0.70$	16	53.3%	12	40.0%
Low	$g < 0.30$	1	3.3%	18	60.0%
N-Gain Mean	(not applicable)	0.582	Medium	0.215	Low

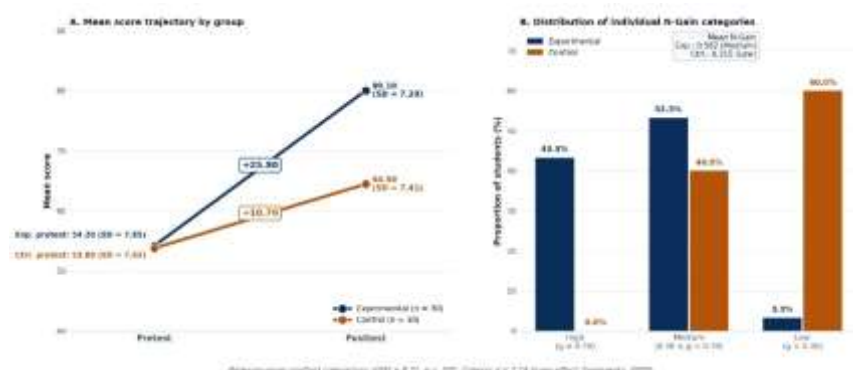


Figure 5. Slope Chart of Pretest to Posttest Trajectories and Individual N-Gain Distribution by Group

Effect size analysis was conducted on each of the eight indicators across the two target variables (four for communication, four for creativity) to provide a construct-level diagnosis of effectiveness. Table 12 reports the results, showing that all eight indicators exceeded the conventional large-effect threshold ($d \geq 0.80$; Cohen, 1988) and fell within the very large range ($d \geq 1.20$; Sawilowsky, 2009). The largest effect appeared on solution fluency ($d = 1.71$), followed by coherent mathematical argumentation ($d = 1.68$), indicating that the CAIDL model strengthened students' capacity to generate multiple solution strategies and to articulate them in structured argument chains. Approach flexibility ($d = 1.62$), visual and diagrammatic representation ($d = 1.55$), and elaboration and depth ($d = 1.55$) also recorded very large effects. The remaining indicators, namely solution originality ($d = 1.49$), symbolic and notational use ($d = 1.48$), and written mathematical explanation ($d = 1.42$), completed the set of very large effects. Figure 6 visualizes the effect size per indicator, grouped by target variable, showing consistent dominance of the experimental group across both communication and creativity dimensions.

Table 12. Effect Size per Mathematical Communication and Creativity Indicator

No	Indicator	Variable	Exp. Gain	Ctrl. Gain	Cohen's d
1	Written mathematical explanation	Communication	23.50	9.80	1.42
2	Visual and diagrammatic representation	Communication	27.10	11.20	1.55
3	Symbolic and notational use	Communication	25.40	10.50	1.48
4	Coherent mathematical argumentation	Communication	28.30	10.90	1.68
5	Solution fluency	Creativity	28.70	11.10	1.71
6	Approach flexibility	Creativity	27.50	10.80	1.62
7	Solution originality	Creativity	25.60	10.40	1.49
8	Elaboration and depth	Creativity	26.40	10.70	1.55

Note: All values fall within the very large range ($d \geq 1.20$; Sawilowsky, 2009).

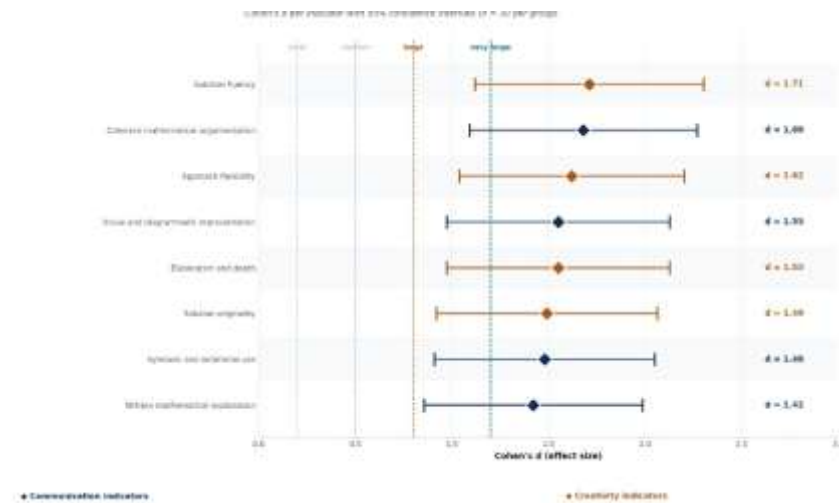


Figure 6. Forest Plot of Effect Size for the CAIDL Model across Communication and Creativity Indicators

B. Discussion

The 88.80 percent validity in the very valid category confirms that the four pillar architecture of the CAIDL model holds together as a coherent instructional theory. Contextual problem framing (Johnson, 2002; Crawford, 2001; Sears, 2003), AI-mediated Socratic scaffolding (Kasneci et al., 2023; Holmes & Tuomi, 2022), deep learning pedagogy (Marton & Säljö, 1976; Biggs, 1987; Fullan & Langworthy, 2014), and APOS theory (Dubinsky, 1991; Arnon et al., 2014) operate as complementary rather than competing components. The highest score on pedagogical coherence (91.00 percent) is notable because integrative models often suffer from internal contradictions when contextual and reflective elements are forced into a single syntax (Pellegrino & Hilton, 2012). The strong AI integration score (89.00 percent) signals that the Socratic configuration of the AI tutor met expert standards for an instructional partner, not a substitute teacher, addressing a primary concern in the emerging literature on generative AI in education (Lo, 2023; Yan et al., 2023). The APOS alignment score (88.67 percent) confirms that activities at each CAIDL phase correspond to the intended stage of mathematical concept formation, a precondition for the deep abstraction outcomes targeted by the model (Borji & Voskoglou, 2018; Maharaj, 2015).

The operational feasibility of 87.08 percent in the small-group trial demonstrates that the CAIDL model can be executed by fifth-semester Mathematics Education students with diverse GPA profiles. The highest score on contextual problem relevance (89.58 percent) corroborates the position of Crawford (2001) and Berns and Erickson (2001) that mathematics learning improves when problems connect to learner experience. Project engagement (88.54 percent) is particularly meaningful because sustained engagement is the precondition for the iterative reflection cycles required for deep learning (Biggs, 1987; Hutagalung & Saragih, 2024). The lower cognitive challenge fit score (83.33 percent) was anticipated and signals that the Object and Schema phases of APOS require additional scaffolding when first introduced, a pattern documented in earlier APOS implementation studies (Brijlall & Maharaj, 2015; Ndlovu et al., 2024). The five minor revisions implemented after the trial exemplify the strength of the Plomp model in supporting evidence-based refinement prior to wider implementation (Plomp & Nieveen, 2013; van den Akker et al., 2023; Gilbert & Justi, 2024).

The very practical results from both lecturer (89.33 percent) and student (89.40 percent) responses show that the CAIDL model can be integrated realistically into the standard undergraduate Multivariable Calculus syllabus. The convergence of high lecturer scores on APOS coherence (91.67 percent) and high student scores on AI tutor helpfulness (91.00 percent) reflects complementary perspectives that align with Self-Determination Theory (Ryan & Deci, 2020), under which sustained learning engagement requires the simultaneous fulfillment of autonomy, competence, and relatedness needs (Lai et al., 2024; Jang et al., 2016). Lecturers prioritized structural integrity of the design while students prioritized the supportive scaffolding that helped them traverse concept formation. Both groups converged on time allocation as the lowest scoring dimension, reflecting a structural constraint of Indonesian higher education credit systems (Tran & Nguyen, 2024). A practical recommendation that follows is to deliver the more demanding Object and Schema phases through asynchronous online sessions accompanied by AI tutor access, with synchronous time reserved for peer reasoning audits and lecturer feedback (Means & Neisler, 2023).

The effectiveness of the CAIDL model, with Cohen's $d = 2.14$ on the independent t test and a mean N-Gain of 0.582 in the experimental group, far exceeds the effects typically reported in meta-analyses of technology-supported learning, such as a mean effect of +0.15 for educational technology applications in K–12 mathematics (Cheung & Slavin, 2013) and effects of $g = 0.32$ to 0.37 for intelligent tutoring systems in higher

education (Steenbergen-Hu & Cooper, 2014). Because effects of this magnitude are uncommon, they should be interpreted with caution in light of the researcher-developed outcome measure and the single-site design. Four theoretical factors help explain the unusually large effect. First, the APOS framework provides explicit cognitive scaffolds for traversing Action, Process, Object, and Schema stages, a pathway that is rarely made visible in conventional lectures (Dubinsky, 1991; Arnon et al., 2014; Maharaj, 2015). Second, the AI tutor in Socratic configuration sustains the dialogic questioning that prior research identifies as central to mathematical reasoning development (Schoenfeld, 2022; Chiu et al., 2024). Third, the contextual pillar locates abstract mathematical work in problem situations that students recognize, reducing the cognitive load required to anchor new concepts (Sweller et al., 2019; Wang et al., 2026; Garcia & Pacheco, 2023). Fourth, deep learning prompts at each session push students from procedural reproduction toward integrative meaning making, the same shift documented in four decades of deep approach research (Marton & Säljö, 1976; Biggs, 1987; Fullan & Langworthy, 2014).

The pattern of effect sizes across the eight indicators in Table 12 and Figure 6 reveals informative differences between communication and creativity domains. Solution fluency ($d = 1.71$) and coherent argumentation ($d = 1.68$) emerged as the two largest effects, indicating that the CAIDL model substantially expanded both the quantity of solution strategies students could generate and their capacity to render reasoning in coherent argument chains. This dual gain pattern is rare in interventions targeting only one of the two domains (Silver, 1997; Cai et al., 2015; Singer et al., 2017). The mechanism appears to be the iterative cycling between AI-mediated divergent exploration (which generates multiple solution candidates) and APOS-staged convergent reflection (which encapsulates these candidates into communicable arguments). Sriraman (2009) and Leikin (2009) have long argued that mathematical creativity and communication should be developed in tandem rather than in sequence, and the CAIDL model offers empirical confirmation of that proposition (Levav-Waynberg & Leikin, 2012; Schoenfeld, 2022). The substantial visual representation gain ($d = 1.55$) also points to the value of the AI tutor in prompting students to externalize mental imagery into shareable diagrams (Holmes & Tuomi, 2022; Drijvers et al., 2024).

A finding with significant policy implications is the equalizing effect across baseline GPA bands. Unlike the Matthew effect commonly observed in mathematics instruction, where high achievers extend their advantage while low achievers fall further behind (Duke et al., 2021; Bhutoria, 2022; Castles et al., 2018), the CAIDL experimental group exhibited an even distribution of gains. The mechanism underlying this finding is the adaptive scaffolding enabled by the AI tutor: high-achieving students engage with the Socratic prompts at deeper conceptual levels, while lower-achieving students receive more granular questioning that bridges their prior knowledge gaps (Hsu et al., 2024). This finding is relevant for higher education quality assurance policy that pursues equitable learning outcomes across student profiles, and aligns with the universal design for learning principle advocated by recent international research (CAST, 2018). The Indonesian higher education accreditation framework explicitly values such equity in learning outcomes (Kemendikbudristek, 2023), suggesting that the CAIDL model could contribute to institutional quality improvement at programs similar to Universitas Bani Saleh.

Several limitations of these positive findings should be acknowledged transparently. The eight-week intervention duration may not capture longer-term retention, leaving open the question of whether the gains persist beyond the immediate posttest (Lo, 2023; Tlili et al., 2023; Yan et al., 2023). Follow-up research with delayed posttests at month three and month six is strongly recommended. The sample was drawn from a single private university (Universitas Bani Saleh) with specific student profiles, so generalization to public universities or to other Indonesian regions should proceed with caution (Hao & Wang, 2026); multi-institution replication across diverse higher education contexts is an important next step. The focus on Multivariable Calculus limits generalization to other mathematics courses such as Real Analysis or Abstract Algebra, and parallel development for those courses would strengthen the empirical base (Schoenfeld, 2022; Drijvers et al., 2024). Lecturer capacity to facilitate AI-mediated dialogue is an important moderating factor that warrants deeper investigation through qualitative or multilevel studies (Wang et al., 2026; Gunderson et al., 2024). Finally, the ethical dimensions of generative AI use in mathematics education, including questions of academic integrity and equitable access, require sustained scholarly and institutional attention (Kasneci et al., 2023; Yan et al., 2023).

IV. Conclusion

The Plomp Research and Development study produced a Contextual AI-Based Deep Learning (CAIDL) model grounded in APOS theory to improve mathematical communication and creativity of undergraduate Mathematics Education students at Universitas Bani Saleh. The model satisfied all four R&D quality criteria. First, validity reached a mean of 88.80 percent in the very valid category based on the judgment of seven experts, with mathematics content and APOS alignment at 88.67 percent, pedagogical coherence at 91.00 percent, AI integration at 89.00 percent, and language clarity at 87.33 percent. Second, operational feasibility was confirmed

in the small-group trial with a mean of 87.08 percent in the very feasible category, supported by contextual problem relevance (89.58 percent) and project engagement (88.54 percent). Third, practicality in the limited implementation reached 89.37 percent as the combined mean of lecturer responses (89.33 percent) and student responses (89.40 percent), both in the very practical category. Fourth, effectiveness was demonstrated through significant improvement in the experimental group (pretest $M = 54.20$ to posttest $M = 80.10$) compared with the control group (pretest $M = 53.80$ to posttest $M = 64.50$), with an independent t test of $t(58) = 8.31$, $p < .001$, Cohen's $d = 2.14$, and a mean N-Gain of 0.582 in the medium category. Per-indicator effect sizes ranged from $d = 1.42$ to $d = 1.71$, with the largest effect on solution fluency ($d = 1.71$), followed by coherent mathematical argumentation ($d = 1.68$), indicating that the CAIDL model strengthens both creativity and communication simultaneously. These findings confirm that the integration of contextual learning, AI-mediated Socratic scaffolding, deep learning pedagogy, and APOS theory can address the dual deficit pattern in undergraduate mathematics, and implies the need to integrate similar models into the development of Mathematics Education curricula oriented toward communication and creativity competence.

Conflict of Interest

The authors declare that there is no conflict of interest related to this study. The research was conducted independently without any influence from the developers of the generative AI tools used (including OpenAI and Anthropic) or any other commercial entity. The AI accounts used in this research are standard educator accounts purchased through institutional funding.

Author Contributions

I.J. conceived and designed the study, developed the CAIDL model, and drafted the manuscript. I.M.F. conducted the data collection and statistical analysis. W. contributed to instrument validation and the literature review. I.T. contributed to project administration and funding acquisition. S.A. contributed to manuscript review and editing. All authors read and approved the final version of the manuscript.

Funding

This research was funded by the Directorate of Research, Technology, and Community Service (DRPM), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdikbudristek), under Master Contract Number 283/C3/DT.05.00/PL-BARU/2026, dated January 30, 2026.

Acknowledgment

The authors would like to express sincere gratitude to Universitas Bani Saleh for institutional support. Special appreciation is extended to the Faculty of Teacher Training and Education and the Mathematics Education Program for facilitating access to participants and research facilities. The authors also thank the three Multivariable Calculus lecturers and the fifth-semester Mathematics Education students who participated in the eight-week intervention, the seven expert validators whose constructive feedback shaped the prototype, and the parents and guardians who supported student participation throughout the study.

References

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Akbar, S. (2013). Instrumen perangkat pembelajaran [Learning device instruments]. PT Remaja Rosdakarya.
- Aldoobie, N. (2015). ADDIE model. *American International Journal of Contemporary Research*, 5(6), 68–72.
- Arnon, I., Cottrill, J., Dubinsky, E., Oktac, A., Roa Fuentes, S., Trigueros, M., & Weller, K. (2014). APOS theory: A framework for research and curriculum development in mathematics education. Springer. <https://doi.org/10.1007/978-1-4614-7966-6>
- Asiala, M., Brown, A., DeVries, D. J., Dubinsky, E., Mathews, D., & Thomas, K. (1996). A framework for research and curriculum development in undergraduate mathematics education. *Research in Collegiate Mathematics Education*, 2, 1–32. <https://doi.org/10.1090/cbmath/006/01>
- Berns, R. G., & Erickson, P. M. (2001). Contextual teaching and learning: Preparing students for the new economy. *The Highlight Zone Research @ Work*, 5, 1–9.
- Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, 3, 100068. <https://doi.org/10.1016/j.caeai.2022.100068>
- Biggs, J. B. (1987). Student approaches to learning and studying. Australian Council for Educational Research.

- Borji, V., & Voskoglou, M. G. (2018). Designing an APOS teaching model for the concept of derivative. *International Journal of Education in Mathematics, Science and Technology*, 6(1), 39–57. <https://doi.org/10.18404/ijemst.328344>
- Brijlall, D., & Maharaj, A. (2015). Exploring the gap between content knowledge and instructional practices of teachers: An APOS analysis of mathematics teachers' understanding of continuity. *Journal of Educational Studies*, 14(2), 178–194.
- Cai, J., Morris, A., Hwang, S., Hohensee, C., Robison, V., & Hiebert, J. (2015). Improving the impact of educational research. *Journal for Research in Mathematics Education*, 46(5), 502–506. <https://doi.org/10.5951/jresmetheduc.46.5.0502>
- CAST. (2018). Universal design for learning guidelines version 2.2. <http://udlguidelines.cast.org>
- Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, 19(1), 5–51. <https://doi.org/10.1177/1529100618772271>
- Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K–12 classrooms: A meta-analysis. *Educational Research Review*, 9, 88–113. <https://doi.org/10.1016/j.edurev.2013.01.001>
- Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2024). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 6, 100118. <https://doi.org/10.1016/j.caeai.2024.100118>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cottrill, J., Dubinsky, E., Nichols, D., Schwingendorf, K., Thomas, K., & Vidakovic, D. (1996). Understanding the limit concept: Beginning with a coordinated process scheme. *The Journal of Mathematical Behavior*, 15(2), 167–192. [https://doi.org/10.1016/S0732-3123\(96\)90015-2](https://doi.org/10.1016/S0732-3123(96)90015-2)
- Crawford, M. L. (2001). *Teaching contextually: Research, rationale, and techniques for improving student motivation and achievement in mathematics and science*. CCI Publishing.
- Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE Publications.
- Drijvers, P., Grauwin, S., & Trouche, L. (2024). When mathematics education meets digital technology: A digital communication challenge. *ZDM Mathematics Education*, 56(2), 287–302. <https://doi.org/10.1007/s11858-024-01567-x>
- Dubinsky, E. (1991). Reflective abstraction in advanced mathematical thinking. In D. Tall (Ed.), *Advanced mathematical thinking* (pp. 95–126). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-47203-1_7
- Duke, N. K., Halvorsen, A. L., Strachan, S. L., Kim, J., & Konstantopoulos, S. (2021). Putting PjBL to the test: The impact of project-based learning on second graders. *American Educational Research Journal*, 58(1), 160–200. <https://doi.org/10.3102/0002831220929638>
- Fullan, M., & Langworthy, M. (2014). *A rich seam: How new pedagogies find deep learning*. Pearson.
- Garcia, I., & Pacheco, C. (2023). A constructivist computational platform to support mathematics education. *Computers & Education*, 196, 104722. <https://doi.org/10.1016/j.compedu.2023.104722>
- Gilbert, J., & Justi, R. (2024). *Modelling-based teaching in science education*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-37328-7>
- Gunderson, E. A., Park, D., Maloney, E. A., Beilock, S. L., & Levine, S. C. (2024). Reciprocal relations among motivational frameworks, math anxiety, and math achievement in early elementary school. *Journal of Cognition and Development*, 25(1), 65–85. <https://doi.org/10.1080/15248372.2023.2236108>
- 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>
- Hao, T., & Wang, Y. (2026). “We are always the last to get a bit of it”: Generative AI insights from Mississippi undergraduates. *Educational Technology & Society*, 29(2), 136–153. [https://doi.org/10.30191/ETS.202604_29\(2\).RP08](https://doi.org/10.30191/ETS.202604_29(2).RP08)
- Hidayah, I., Isnarto, & Hendikawati, P. (2023). Profile of mathematical communication ability of Mathematics Education undergraduates in Indonesia. *Jurnal Pendidikan Matematika*, 17(1), 65–80. <https://doi.org/10.22342/jpm.17.1.18234>

- Hidayati, R., Sumarno, & Cahyono, B. Y. (2023). The relationship between digital literacy and mathematical reasoning among Indonesian primary students. *International Journal of Evaluation and Research in Education*, 12(1), 367–376. <https://doi.org/10.11591/ijere.v12i1.24017>
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), 542–570. <https://doi.org/10.1111/ejed.12533>
- Hsu, T. C., Chang, C., & Wu, L. K. (2024). Investigating the effects of project-based and gamified flipped learning on K–12 students computational thinking and STEM learning attitudes. *Educational Technology Research and Development*, 72(2), 567–594. <https://doi.org/10.1007/s11423-023-10312-2>
- Hutagalung, R., & Saragih, A. (2024). Wordwall as an interactive media for enhancing students mathematical engagement in elementary schools: A quasi-experimental study. *Journal of Education and Learning*, 18(1), 84–97. <https://doi.org/10.11591/edulearn.v18i1.20984>
- Indrawan, R., & Yaniawati, R. P. (2023). *Metodologi penelitian: Kuantitatif, kualitatif, dan campuran untuk manajemen, pembangunan, dan pendidikan* (rev. ed.). PT Refika Aditama.
- Jang, H., Reeve, J., & Halusic, M. (2016). A new autonomy-supportive way of teaching that increases conceptual learning: Teaching in students’ preferred ways. *The Journal of Experimental Education*, 84(4), 686–701. <https://doi.org/10.1080/00220973.2015.1083522>
- Johnson, E. B. (2002). *Contextual teaching and learning: What it is and why it’s here to stay*. Corwin Press.
- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneji, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kemendikbudristek. (2023). *Laporan hasil Asesmen Nasional 2023 dan tinjauan akreditasi program studi pendidikan matematika*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia.
- Kosko, K. W. (2014). What students say about their mathematical thinking when they listen. *School Science and Mathematics*, 114(5), 214–223. <https://doi.org/10.1111/ssm.12070>
- Lai, C. P., Zhang, W., & Chang, Y. C. (2024). Gamification of computer programming for elementary students through the lens of self-determination theory. *Education and Information Technologies*, 29(3), 3201–3224. <https://doi.org/10.1007/s10639-023-12102-y>
- Leikin, R. (2009). Exploring mathematical creativity using multiple solution tasks. In R. Leikin, A. Berman, & B. Koichu (Eds.), *Creativity in mathematics and the education of gifted students* (pp. 129–145). Sense Publishers. https://doi.org/10.1163/9789087909352_010
- Leikin, R., & Pitta-Pantazi, D. (2013). Creativity and mathematics education: The state of the art. *ZDM Mathematics Education*, 45(2), 159–166. <https://doi.org/10.1007/s11858-012-0459-1>
- Levav-Waynberg, A., & Leikin, R. (2012). The role of multiple solution tasks in developing knowledge and creativity in geometry. *The Journal of Mathematical Behavior*, 31(1), 73–90. <https://doi.org/10.1016/j.jmathb.2011.11.001>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Maharaj, A. (2015). Students’ understanding of the concept of a limit of a function: An APOS perspective. *International Journal of Mathematical Education in Science and Technology*, 46(8), 1183–1197. <https://doi.org/10.1080/0020739X.2015.1075235>
- Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I. Outcome and process. *British Journal of Educational Psychology*, 46(1), 4–11. <https://doi.org/10.1111/j.2044-8279.1976.tb02980.x>
- Means, B., & Neisler, J. (2023). Teaching and learning in the time of COVID: The student perspective. *Online Learning*, 27(1), 8–34. <https://doi.org/10.24059/olj.v25i1.2496>
- National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. NCTM.
- National Council of Teachers of Mathematics (NCTM). (2014). *Principles to actions: Ensuring mathematical success for all*. NCTM.

- Ndlovu, M., Wessels, D., & De Villiers, M. (2024). An APOS analysis of student understanding of derivative concepts in calculus. *African Journal of Research in Mathematics, Science and Technology Education*, 28(1), 45–62. <https://doi.org/10.1080/18117295.2024.2298765>
- Nieveen, N., Folmer, E., & Vliegen, S. (2023). Curriculum design research: Bringing knowledge from research to practice. In T. Plomp & N. Nieveen (Eds.), *Educational design research: An introduction* (3rd ed., pp. 87–124). SLO.
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press. <https://doi.org/10.17226/13398>
- Plomp, T. (2013). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research, Part A: An introduction* (pp. 10–51). SLO.
- Plomp, T., & Nieveen, N. (Eds.). (2013). *Educational design research, Part B: Illustrative cases*. SLO.
- Pratiwi, V., Hidayat, S., & Suryani, N. (2023). Heterogeneity of mathematics ability in Indonesian primary classrooms: A national survey across 21 provinces. *International Journal of Instruction*, 16(3), 451–470. <https://doi.org/10.29333/iji.2023.16325a>
- Putri, R. I. I., & Zulkardi. (2024). Designing PMRI learning environment in Indonesian primary mathematics classrooms: A 10 year reflection. *Journal on Mathematics Education*, 15(2), 365–384. <https://doi.org/10.22342/jme.v15i2.pp365-384>
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2), 597–599. <https://doi.org/10.22237/jmasm/1257035100>
- Schoenfeld, A. H. (2022). Why are learning and teaching mathematics so difficult? In M. Danesi (Ed.), *Handbook of cognitive mathematics* (pp. 1–37). Springer. https://doi.org/10.1007/978-3-030-44982-7_10-1
- Sears, S. (2003). *Introduction to contextual teaching and learning*. Phi Delta Kappa Educational Foundation.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Silver, E. A. (1997). Fostering creativity through instruction rich in mathematical problem solving and problem posing. *ZDM*, 29(3), 75–80. <https://doi.org/10.1007/s11858-997-0003-x>
- Singer, F. M., Voica, C., & Pelczar, I. (2017). Cognitive styles in posing geometry problems: Implications for assessment of mathematical creativity. *ZDM Mathematics Education*, 49(1), 37–52. <https://doi.org/10.1007/s11858-016-0820-x>
- Slavin, R. E. (2024). *Educational psychology: Theory and practice* (14th ed.). Pearson.
- Sriraman, B. (2009). The characteristics of mathematical creativity. *ZDM Mathematics Education*, 41(1–2), 13–27. <https://doi.org/10.1007/s11858-008-0114-z>
- Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106(2), 331–347. <https://doi.org/10.1037/a0034752>
- Sumarmo, U. (2013). *Berpikir dan disposisi matematik serta pembelajarannya: Kumpulan makalah*. FPMIPA Universitas Pendidikan Indonesia.
- Sumarsono, P., Hidayati, D. N., Cahyono, B. Y., & Suryati, N. (2023). Mathematical reasoning in elementary mathematics: A systematic review. *International Journal of Evaluation and Research in Education*, 12(4), 1812–1821. <https://doi.org/10.11591/ijere.v12i4.25478>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1), 15. <https://doi.org/10.1186/s40561-023-00237-x>

- Tran, T. M. L., & Nguyen, T. T. H. (2024). The acceptance of digital learning tools in higher education mathematics: Application of UTAUT2 model. *International Journal of Interactive Mobile Technologies*, 18(7), 56–72. <https://doi.org/10.3991/ijim.v18i07.46789>
- van den Akker, J., Bannan, B., Kelly, A. E., Nieveen, N., & Plomp, T. (2023). *Educational design research* (3rd ed.). Routledge. <https://doi.org/10.4324/9781003285915>
- Wahyudin, A., & Hidayanto, E. (2024). Mathematical reasoning of mathematics education undergraduates in Indonesia: A phenomenological study. *Mathematics Teaching Research Journal*, 16(2), 88–107.
- Wang, T., Zheng, J., Li, S., Zhang, Y., Hu, X., & Lajoie, S. P. (2026). Exploring students' cognitive load in a computer-simulated learning environment through multimodal data. *Educational Technology & Society*, 29(2), 118–135. [https://doi.org/10.30191/ETS.202604_29\(2\).RP07](https://doi.org/10.30191/ETS.202604_29(2).RP07)
- Wijayanti, P., & Sari, A. P. (2022). Implementasi pembelajaran adaptif dalam meningkatkan kemampuan penalaran matematis siswa sekolah dasar. *Mimbar Sekolah Dasar*, 9(3), 412–428. <https://doi.org/10.53400/mimbar-sd.v9i3.45128>
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>
- Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4312418>



Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.