# Learning Changes Following a Structured Metacognitive Intervention among Lower-Performing Undergraduate Students in Calculus Limits

Xiaoxia Liu[1], Sian Hoon Teoh[2,*], Geethanjali Narayanan[3], Walton Wider[4], Xiaoai Zeng[5]

[1,2,3]*Faculty of Education, Universiti Teknologi MARA, Puncak Alam Campus, 42300 Puncak Alam, Selangor, Malaysia*

[1]*School of Statistics and Data Science, Hebei Finance University, No. 3188 Hengxiang North Street, Baoding 071051, Hebei Province, China*

[4]*Faculty of Business and Communications, INTI International University, Persiaran Perdana BBN, Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia*

[4]*Faculty of Economics and Business, Tarumanagara University, Jl. Letjen S. Parman No.1, RT.6/RW.16, Tomang, Kec. Grogol Petamburan, Kota Jakarta Barat, Daerah Khusus, 11440 Jakarta, Indonesia*

[5]*Faculty of Psychology, Shinawatra University, 99, Bang Toei, Sam Khok District, 12160 Pathum Thani, Thailand*

(Received: March 29, 2026; Revised: June 1, 2026; Accepted: September 1, 2026; Available online: October 29, 2026)

**Abstract**

This study examined within-student learning changes during a six-week structured metacognitive intervention among 46 lower-performing undergraduate students studying calculus limits. Participants were identified using a pragmatic, cohort-relative bottom 30% screening rule and completed content-matched 100-point achievement forms in a one-group pretest-posttest design. Mean achievement increased from 34.70 (SD = 19.56) to 59.54 (SD = 18.12), representing a mean gain of 24.85 points, $t(45) = 8.27$, $p < .001$, $dz = 1.22$. Absolute point gains differed significantly across low-, medium-, and high-difficulty categories, $F(2, 90) = 13.87$, $p < .001$, partial $\eta^2 = .24$. Medium-difficulty absolute gains exceeded low- and high-difficulty gains, whereas low- and high-difficulty gains did not differ significantly. Because response-format score maxima differed between forms, scores were normalized to percentages of the available maximum. Percentage-point gains differed significantly across multiple-choice, true-false, and short-answer formats after Greenhouse-Geisser correction, $F(1.74, 78.18) = 8.07$, $p = .001$, partial $\eta^2 = .15$. Short-answer gains exceeded those for multiple-choice and true-false formats, whereas the latter two did not differ significantly. Completion time increased significantly by 16.87 minutes, whereas the score-time ratio showed no statistically significant change. These findings characterize within-student performance changes rather than causal treatment effects, as maturation, repeated testing, concurrent instruction, instructor-related influences, and regression to the mean remain plausible alternative explanations. Planning, monitoring, evaluation, cognitive load, and implementation fidelity were not directly measured. Controlled studies incorporating direct process measures, prospective fidelity assessment, and psychometrically calibrated achievement forms are needed to test efficacy and proposed mechanisms.

*Keywords:* Metacognitive Scaffolding, Calculus Limits, Lower-Performing Students, Self-Regulated Learning, Mathematics Achievement, Time-On-Task, Education Quality

## 1. Introduction

Calculus is a foundational area of advanced mathematical learning in higher education and supports analytical reasoning and mathematical modelling in quantitatively intensive disciplines such as physics and engineering [1], [2]. Yet many undergraduates struggle persistently with calculus, particularly with conceptually demanding topics such as limits that require abstraction and coordination across symbolic, graphical, and numerical representations [3], [4], [5]. Research on students' conceptions of limits underscores the need for support that helps them engage deliberately with the topic's conceptual and procedural demands [4], [5].

These challenges may be especially consequential for students entering calculus with lower baseline performance. Such learners may struggle to integrate conceptual understanding with procedural reasoning and coordinate multiple forms of mathematical representation [4], [5]. Prior research has consistently associated metacognitive awareness and

***Corresponding author: Sian Hoon Teoh (teohsian@uitm.edu.my)**
**DOI: https://doi.org/10.47738/jads.v7i4.1481**
**This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).**
**© Authors retain all copyrights**

regulation with mathematics achievement, reasoning, modelling competence, and other dimensions of mathematical performance [6], [7], [8], [9], [10]. Weaker metacognitive engagement may therefore be associated with less effective planning, monitoring, evaluation, and persistence during complex problem solving [11], [12], [13]. These associations support instructional approaches that make strategic problem-solving processes more explicit, although they do not imply that all lower-performing students necessarily have deficits in metacognitive regulation.

Metacognition refers to the awareness and regulation of cognition through processes such as planning, monitoring, and evaluation [14], [15], [16], [17]. In mathematics learning, teacher modelling, structured self-questioning, guided reflection, and related forms of scaffolding can support students' strategic engagement with complex problems and have been associated with stronger mathematical understanding and achievement [18], [19], [20], [21], [22]. More broadly, research on self-regulated learning indicates that explicit strategy instruction and structured scaffolding can help learners develop greater control over their learning processes [23]. In the present study, these regulatory processes provide the theoretical rationale for the instructional design rather than directly measured outcomes.

Three more specific gaps motivate the study. First, although metacognitive interventions have been examined across mathematics education, evidence remains comparatively limited regarding structured metacognitive scaffolding for lower-performing undergraduates specifically learning calculus limits. Emerging research has begun to examine metacognitive scaffolding in learning limit concepts [4], [5], but more evidence is needed on how performance changes unfold across different assessment dimensions. Second, relatively little attention has been given to whether pretest-posttest performance changes vary according to task difficulty and response format within the same cohort. Response format matters because constructed-response tasks require learners to generate solutions, whereas multiple-choice and true-false formats involve different recognition and response processes [24]. Third, time-based measures are difficult to interpret because longer completion time may reflect productive checking and persistence, but may also indicate uncertainty, cognitive overload, or inefficient strategy use [25], [26]. Examining these dimensions together can therefore provide a more differentiated account of learning change than overall achievement scores alone.

Accordingly, this study examines within-student changes in calculus-limits achievement, task-difficulty performance, response-format performance, completion time, and score-time ratio during a six-week structured intervention. The objective is to characterize the pattern of change across these outcomes, not to estimate a causal treatment effect or directly measure metacognitive development.

## 2. Literature Review

### 2.1. Theoretical Underpinning

Metacognitive regulation comprises planning before action, monitoring during task performance, and evaluation following task completion [14], [15]. Within self-regulated learning models, these processes operate within iterative cycles of goal setting, strategy selection, performance monitoring, feedback, and reflection [16], [17]. In calculus problem solving, these regulatory processes are theoretically relevant because students must select appropriate strategies, coordinate symbolic, graphical, and numerical representations, monitor intermediate steps, and evaluate the plausibility and accuracy of their solutions.

These principles informed the intervention's structure. In particular, the instructional activities incorporated teacher modelling, structured self-questioning, strategic checking, guided reflection, and progressive reduction of external prompts. The metacognitive-questioning framework described in [20] further informed this design through four complementary domains: comprehension, connection, strategy, and reflection. These domains correspond to understanding problem requirements, linking relevant concepts and representations, selecting and monitoring solution strategies, and evaluating completed work. Similar approaches have subsequently been used to scaffold metacognitive engagement during mathematical problem solving [19], [20]. In the present study, however, these constructs informed the intervention design and proposed mechanism rather than serving as directly measured study outcomes.

### 2.2. Metacognition in Mathematics Learning

Research in mathematics education consistently links metacognitive regulation with stronger mathematical reasoning, problem solving, and achievement [6], [7]. Students who engage more effectively in strategic monitoring and reflection

tend to demonstrate deeper conceptual understanding, more adaptive problem-solving approaches, stronger mathematical modelling competence, and higher academic performance [8], [9], [10], [21]. Metacognitive problem solving has also been associated with students' cognitive, behavioral, and affective engagement during complex learning activities [13].

These processes are especially relevant in calculus, where students must coordinate abstract concepts with symbolic, graphical, and numerical representations. Limits are particularly demanding because successful problem solving often requires students to connect formal definitions, graphical interpretations, numerical patterns, and symbolic procedures. Research on students' conceptions and temporal reasoning in limits highlights the conceptual demands that such scaffolding is intended to address [4], [5]. For lower-performing learners, difficulty managing such demands may contribute to reliance on surface-level procedures, premature disengagement, or ineffective strategy use as tasks become more complex [2], [12], [21].

The association between metacognitive regulation and mathematics achievement supports instructional support that makes strategic problem solving more explicit [6], [7]. However, such associations do not, by themselves, show that an instructional intervention changes students' underlying regulatory processes. This distinction matters when metacognition is inferred from achievement, reasoning, engagement, or other behavioral outcomes rather than assessed directly through validated inventories, observations, think-aloud protocols, or related process measures [12]. Accordingly, improvements in mathematical performance should not automatically be interpreted as evidence of improved metacognitive regulation.

## 2.3. Metacognitive Interventions and Academic Achievement

Metacognitive scaffolding in mathematics commonly incorporates teacher modelling, structured self-questioning, guided error analysis, reflection, feedback, and opportunities for students to evaluate their own solution processes. These approaches are intended to make otherwise implicit problem-solving processes more explicit and to support students in selecting, monitoring, and evaluating strategies during mathematical tasks [18], [19], [20], [22]. Research on self-regulated learning similarly suggests that explicit strategy instruction and structured instructional support can help learners develop more deliberate approaches to planning, monitoring, and evaluating their learning [23].

Previous intervention studies and research syntheses have reported positive associations between metacognitive instructional support and mathematics achievement [18], [19], [21]. Metacognitive questioning embedded within mathematical tasks has also been used to encourage learners to interpret task requirements, establish conceptual connections, select appropriate strategies, and reflect on completed solutions [20], [22]. More specifically, documented student difficulties with conceptions and temporal order in limits suggest that structured strategic support may be relevant to conceptually demanding calculus topics [4], [5].

Such support may be especially relevant for learners who require greater instructional structure when confronting unfamiliar or complex mathematical problems. However, achievement gains following metacognitive interventions should be distinguished from demonstrated changes in metacognitive regulation. Without direct process measures, performance improvements cannot establish whether students became more effective in planning, monitoring, evaluating, or regulating their learning. The present study therefore extends prior performance-focused research to the learning of calculus limits while maintaining a clear distinction between observed achievement changes and theoretically proposed regulatory processes.

## 2.4. Task Difficulty, Response Format, and Score-Time Measures

Task difficulty may moderate the performance demands under which scaffolding is used. Moderately demanding tasks can require students to select and monitor strategies while remaining within manageable cognitive demands, whereas more difficult tasks may require greater coordination of multiple steps and representations and place heavier demands on working-memory resources [25], [26]. Metacognitive regulation may become particularly important under such conditions because students must monitor their understanding, select among alternative strategies, and decide whether a current approach remains viable as task demands increase [12], [13]. Such an account provides a theoretical basis for examining whether performance changes differ across task-difficulty categories. However, because the present study

did not directly measure cognitive load, mental effort, or strategy use, differences across difficulty levels cannot be interpreted as direct evidence of underlying cognitive-load or metacognitive mechanisms.

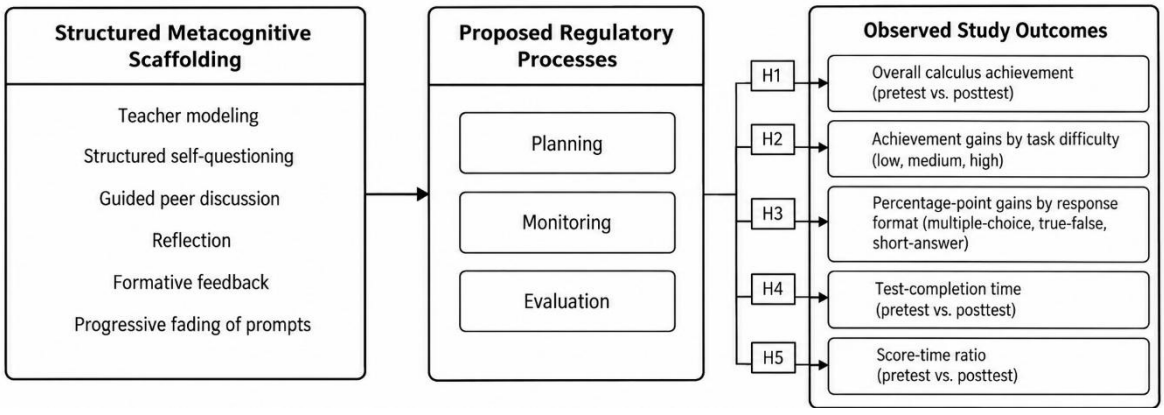
Response format represents another potentially important source of variation in performance. Constructed-response tasks require students to generate solutions and often make intermediate reasoning visible, whereas multiple-choice and true-false formats permit greater reliance on recognition and, to varying degrees, guessing. Research comparing recognition-based assessment formats indicates that multiple-choice and true-false questions may differ in the cognitive and retrieval processes they elicit, making response format relevant when interpreting learning outcomes [24]. Differences in gains across response formats may therefore reflect several factors, including solution-generation demands, recognition processes, score ranges, item discrimination, guessing probabilities, and test composition. Accordingly, comparisons across response formats require careful attention to differences in available score ranges and the comparability of the pretest and posttest forms.

Time-based indicators can complement achievement outcomes but are inherently difficult to interpret. Greater completion time may reflect productive checking, persistence, strategic monitoring, and deliberate processing, but it can also indicate uncertainty, cognitive overload, inefficient strategy selection, or procedural slowness [25], [26]. Metacognitive engagement may also involve students allocating additional time to monitoring and evaluating their responses, yet time expenditure alone cannot determine whether such regulation actually occurred. Similarly, a score-time ratio describes performance relative to time spent but should not automatically be interpreted as a validated measure of learning efficiency or metacognitive regulation.

Taken together, overall achievement, task difficulty, response format, and time-based measures provide complementary perspectives on pretest-posttest performance change among lower-performing students. These dimensions can identify where changes are concentrated and where they are absent, but they do not directly establish changes in planning, monitoring, evaluation, or other metacognitive processes. Direct assessment of these processes would require validated metacognitive measures, behavioral or observational evidence, and more fine-grained indicators of students' strategy use during mathematical problem solving.

## Research Framework and Hypotheses

The proposed research framework links structured metacognitive scaffolding to planning, monitoring, and evaluation as theoretically proposed regulatory processes and, in turn, to the observed study outcomes. Figure 1 presents these hypothesized pathways. They were not empirically tested because the regulatory processes were not directly measured.




Note. Arrows indicate the hypothesized relationships between structured metacognitive scaffolding, proposed regulatory processes, and observed study outcomes. Planning, monitoring, and evaluation were not directly measured; therefore, the proposed regulatory processes represent a theoretical explanation rather than an empirically tested pathway.

**Figure 1.** Proposed research framework.

Based on the theoretical and empirical foundations outlined above, the following hypotheses were proposed:

*H1: Students' overall calculus achievement scores will be significantly higher at posttest than at pretest.*

*H2: Pretest-posttest achievement gains will differ across low-, medium-, and high-difficulty task categories.*

*H3: Pretest-posttest percentage-point gains will differ across multiple-choice, true-false, and short-answer response formats.*

*H4: Students' test-completion time will differ significantly between the pretest and posttest.*

*H5: Students' score-time ratio will differ significantly between the pretest and posttest.*

## 3. Methodology

### 3.1. Research Design

This study employed a one-group pretest-posttest quasi-experimental design to examine within-student changes during a six-week structured intervention embedded within regular calculus instruction. Because the study did not include random assignment or a comparison group, observed pretest-posttest differences cannot be attributed exclusively to the intervention. Potential alternative explanations include maturation, repeated exposure to similar test formats, concurrent classroom instruction, regression toward the mean, routine course progression, and instructor-related influences.

The empirical outcomes were overall achievement, achievement by task difficulty, achievement by response format, test-completion time, and score-time ratio. Planning, monitoring, evaluation, self-regulated learning, cognitive load, mental effort, and engagement were not directly measured and were therefore treated as theoretically proposed processes rather than empirical outcomes.

### 3.2. Participants

Participants were 46 undergraduates enrolled in a compulsory calculus course at Hebei Finance University, China. Screening incorporated prior China National College Entrance Examination performance and a diagnostic assessment administered at the beginning of the semester. The term lower-performing refers to participants' relative standing within the study cohort and should not be interpreted as a standardised or nationally normed classification of mathematical ability.

The bottom 30% threshold was selected before analysis as a pragmatic instructional screening rule rather than as a validated psychometric cut score. It was intended to identify a manageable cohort-relative subgroup for targeted instructional support while retaining sufficient participants for analysis. The threshold was not derived from a standardised benchmark, and alternative cutoffs could have produced different sample compositions and estimates. A cutoff-sensitivity analysis was not possible with the retained dataset. Accordingly, the 30% threshold should not be interpreted as a generalisable definition of mathematical ability.

Participants represented multiple academic majors and included both male and female students. None had received prior formal instruction in metacognitive strategies. All 46 participants completed both assessments, resulting in no attrition between pretest and posttest.

### 3.3. Intervention Procedure

The six-week intervention was embedded within regular instruction on calculus limits and was informed by metacognitive and self-regulated learning frameworks. Each instructional cycle incorporated planning, monitoring, and evaluation activities within mathematical problem solving rather than treating metacognition as a separate instructional component.

During planning, students identified task requirements, relevant prior knowledge, known and unknown information, and candidate solution strategies. During monitoring, they checked intermediate steps, identified inconsistencies, coordinated symbolic, graphical, and numerical representations, and used structured self-questioning to assess the validity of their reasoning. During evaluation, students reviewed solution accuracy and efficiency, considered alternative approaches, and analysed errors to refine subsequent problem solving.

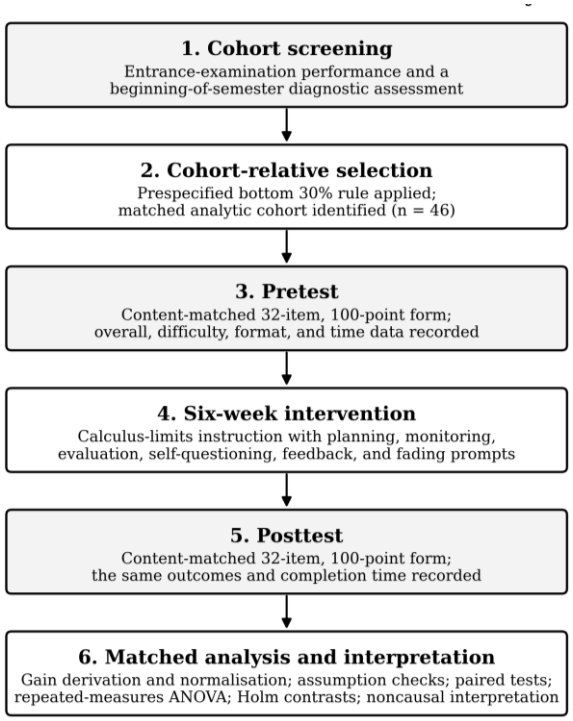
Instructional support included teacher modelling, structured self-questioning, guided peer discussion, reflective learning logs, and formative feedback. Scaffolding was intended to fade progressively from explicit modelling and full prompts in earlier sessions to partial prompts and peer-supported monitoring in the middle phase, and brief self-initiated

 ISSN 2723-6471


checks in later sessions. No independent fidelity observations or session-level adherence records were collected; consequently, the consistency of the intended fading sequence cannot be empirically verified. Reflective logs were used pedagogically and were not systematically coded as research data.

## 3.4. Research Procedure

The same 46 participants completed both assessments. The study sequence comprised cohort screening, selection using the prespecified bottom 30% rule, pretesting, six weeks of instruction, posttesting, matched outcome derivation, and statistical analysis. The planning, monitoring, and evaluation processes informed the intervention but were not directly measured. Figure 2 summarizes participant screening, intervention delivery, outcome derivation, and statistical analysis.




Note. The intervention was embedded in regular instruction. Planning, monitoring, and evaluation informed its design but were not directly measured.

**Figure 2.** Research procedure and statistical analysis flowchart.

## 3.5. Measures and Instruments

Two 32-item calculus-limits achievement forms, each with a maximum score of 100 points, were administered at pretest and posttest. Both forms assessed the same broad content domain and included multiple-choice, true-false, and short-answer responses, although their response-format composition was not identical. The pretest comprised six multiple-choice items worth 12 points, 10 true-false items worth 12 points, and 16 short-answer items worth 76 points. The posttest comprised five multiple-choice items worth 10 points, 10 true-false items worth 12 points, and 17 short-answer items worth 78 points. Maximum scores for the low-, medium-, and high-difficulty categories were 26, 45, and 29 points, respectively.

Achievement was operationalised as the total test score and was further examined by task difficulty and response format. Because the maximum scores assigned to response formats differed between forms, response-format scores were divided by their form-specific maxima and converted to percentages before comparing gain scores. Completion time was recorded in minutes. The score-time ratio was calculated as total achievement score divided by completion time and was treated as an exploratory performance-rate indicator rather than as a validated measure of learning efficiency, engagement, or metacognitive regulation.

Experienced mathematics instructors classified items into low-, medium-, and high-difficulty categories according to conceptual demand, number of procedural steps, and representational complexity. Three senior mathematics educators

reviewed the achievement forms for content coverage and clarity, and this study pilot-tested them with a comparable cohort. Internal consistency was acceptable for both the pretest (Cronbach's α = .82) and posttest (Cronbach's α = .85).

Formal content-validity indices, inter-rater agreement for difficulty classification, item-discrimination statistics, cut-score validation, Item Response Theory or Rasch calibration, measurement-invariance evidence, and full parallel-form equivalence were unavailable. The forms are therefore described as content-matched rather than psychometrically equivalent, and response-format findings are interpreted accordingly. The retained study records also did not include detailed item-level psychometric information for the diagnostic screening assessment.

The study did not collect a validated metacognitive questionnaire, think-aloud protocol, observational coding, mental-effort rating, or item-level response-latency measure. Accordingly, the empirical analyses evaluate performance and time outcomes rather than changes in metacognitive regulation. Table 1 summarizes the study measures, analytical dimensions, and operational definitions.

**Table 1.** Overview of Measures and Analytical Dimensions

| Measure | Dimension | Operationalization |
|---|---|---|
| Achievement | Overall score | Total calculus-limits test score, 0-100 |
| Task difficulty | Low, medium, high | Item categories based on conceptual demand, procedural steps, and representational complexity |
| Response format | Multiple-choice, true-false, short-answer | Scores by response format; percentage-of-maximum scores used for between-format gain comparisons |
| Time-on-task | Completion time | Total minutes required to complete the assessment |
| Score-time ratio | Performance rate | Achievement score divided by completion time; exploratory indicator |

## 3.6. Data Transformation and Statistical Formulation

Let i = 1, …, n denote participants, t denote the assessment occasion, d denote task difficulty, and f denote response format. All change scores were oriented as posttest minus pretest, so positive values indicated an increase.

$$P_{ift} = 100\left(\frac{X_{ift}}{M_{ft}}\right) \tag{1}$$

$$\Delta P_{if} = P_{if,\text{post}} - P_{if,\text{pre}} \tag{2}$$

$$\Delta Y_i = Y_{i,\text{post}} - Y_{i,\text{pre}} \tag{3a}$$

$$\Delta Y_{id} = Y_{id,\text{post}} - Y_{id,\text{pre}} \tag{3b}$$

$$R_{id} = 100\left(\frac{\Delta Y_{id}}{M_d}\right) \tag{4}$$

$$Q_{it} = \frac{Y_{it}}{T_{it}}, \qquad \Delta Q_i = Q_{i,\text{post}} - Q_{i,\text{pre}} \tag{5}$$

$$t = \frac{\bar{\Delta}}{s_\Delta/\sqrt{n}}, \qquad d_z = \frac{\bar{\Delta}}{s_\Delta} = \frac{t}{\sqrt{n}} \tag{6}$$

$$95\%\, CI = \bar{\Delta} \pm t_{.975,n-1}\left(\frac{s_\Delta}{\sqrt{n}}\right) \tag{7}$$

In (1), X is a raw response-format score, and M is its form-specific maximum; (2) gives the percentage-point gain. Equations (3a), (3b), and (4) define overall, difficulty-specific, and descriptive relative gains. In (5), T is completion time in minutes, and the reported mean score-time ratios are means of participant-level ratios. In (6) and (7), $\bar{\Delta}$ and $s\Delta$ are the mean and standard deviation of paired differences.

## 3.7. Data Analysis

All change scores were calculated as posttest minus pretest, using a consistent sign convention. Descriptive statistics were calculated for each outcome, and Shapiro-Wilk tests were used to assess the distribution of paired differences. Total-score and task-difficulty differences did not depart significantly from normality. Completion-time and score-time-ratio differences departed from normality; therefore, this study interpreted the corresponding paired-samples t tests cautiously.

This study evaluated H1, H4, and H5 using paired-samples t tests. Cohen's dz was calculated as the mean paired difference divided by the standard deviation of the paired differences, and 95% confidence intervals were reported for mean differences. For H2, this study compared individual posttest-minus-pretest raw gain scores across low-, medium, and high-difficulty categories using a one-way repeated-measures analysis of variance. For H3, response-format scores were first converted to percentages of their form-specific maxima, after which individual percentage-point gains were compared across multiple-choice, true-false, and short-answer formats using repeated-measures analysis of variance. This study used Mauchly's test to assess sphericity, and applied Greenhouse-Geisser corrections when the assumption was violated. This study followed significant omnibus tests with paired contrasts using Holm-adjusted p-values and unadjusted 95% confidence intervals. Category-specific paired-samples t tests were retained as complementary exploratory analyses.

A sensitivity analysis indicated that, with 46 paired observations and a two-sided α of 0.05, the paired-samples comparisons had approximately 80% power to detect dz ≈ 0.42. Smaller paired effects could therefore have gone undetected. This estimate pertains to the paired comparisons and should not be interpreted as establishing equivalent sensitivity for all repeated-measures contrasts.

## 4. Results and Discussion

Results are reported for overall achievement, task-difficulty gains, normalised response-format gains, completion time, and score-time ratio.

### 4.1. Overall Achievement

Overall calculus achievement was higher at posttest than at pretest. Mean achievement increased from 34.70 (SD = 19.56) to 59.54 (SD = 18.12), corresponding to a mean within-student gain of 24.85 points. The difference was statistically significant, $t(45) = 8.27$, $p < .001$, 95% CI [18.80, 30.90], with a large paired effect size ($dz = 1.22$). This result was consistent with H1. It represents a within-student pretest-posttest difference and does not establish that the intervention was the sole cause of the observed gain. Table 2 presents the descriptive statistics for overall achievement at both assessment occasions.

**Table 2.** Descriptive Statistics for Overall Achievement (N = 46)

| Test occasion | M | SD | Min | Max |
|---|---|---|---|---|
| Pretest | 34.70 | 19.56 | 8 | 76 |
| Posttest | 59.54 | 18.12 | 16 | 90 |

Note. M = mean; SD = standard deviation.

### 4.2. Achievement by Task Difficulty

To evaluate H2, this study compared posttest-minus-pretest raw gains across low-, medium-, and high-difficulty categories. Mauchly's test was not significant, $W = 0.93$, $\chi^2(2) = 3.04$, $p = 0.219$, indicating that the sphericity assumption was satisfied. The repeated-measures analysis showed a significant difference in mean gain across difficulty categories, $F(2, 90) = 13.87$, $p < 0.001$, partial $\eta^2 = .24$. All three categories showed significant within-student gains. Low-difficulty scores increased by 7.46 points, $t(45) = 8.14$, $p < 0.001$, $dz = 1.20$, 95% CI [5.61, 9.30]. Medium-difficulty scores increased by 12.09 points, $t(45) = 7.39$, $p < 0.001$, $dz = 1.09$, 95% CI [8.79, 15.38]. High-difficulty scores increased by 5.30 points, $t(45) = 4.78$, $p < 0.001$, $dz = 0.70$, 95% CI [3.07, 7.54].

Holm-adjusted contrasts indicated that the medium-difficulty absolute gain exceeded the low-difficulty gain by 4.63 points, $t(45) = 3.24$, adjusted $p = .004$, 95% CI [1.75, 7.51], and exceeded the high-difficulty gain by 6.78 points, $t(45) = 4.97$, adjusted $p < 0.001$, 95% CI [4.03, 9.53]. The low- and high-difficulty gains did not differ significantly, adjusted $p = 0.065$. These findings were consistent with H2. The absolute gains should nevertheless be interpreted in relation to the available score ranges. Relative to the maximum score within each category, gains represented approximately 28.7% for low-difficulty items, 26.9% for medium-difficulty items, and 18.3% for high-difficulty items. Thus, although the medium-difficulty category showed the largest absolute gain, the low-difficulty category showed a slightly larger gain relative to its available maximum. Table 3 summarizes the category-specific descriptive statistics, paired changes, and omnibus repeated-measures result.

ISSN 2723-6471


**Table 3.** Within-Student Achievement Changes by Task Difficulty (N = 46)

| Difficulty | Pretest M (SD) | Posttest M (SD) | Mean gain | t(45) | p | dz | Gain as % of category maximum |
|---|---|---|---|---|---|---|---|
| Low | 10.24 (5.37) | 17.70 (4.41) | 7.46 | 8.14 | < .001 | 1.20 | 28.7% |
| Medium | 14.76 (9.05) | 26.85 (9.63) | 12.09 | 7.39 | < .001 | 1.09 | 26.9% |
| High | 9.70 (6.63) | 15.00 (7.27) | 5.30 | 4.78 | < .001 | 0.70 | 18.3% |

Note. Gains equal posttest minus pretest. Category maxima were 26, 45, and 29 points for low-, medium-, and high-difficulty items, respectively. Mauchly's W = .93, p = .219; omnibus F(2, 90) = 13.87, p < .001, partial $\eta^2$ = .24. Holm-adjusted contrasts: medium > low, adjusted p = .004; medium > high, adjusted p < .001; low versus high, adjusted p = .065. Gains were calculated from unrounded participant-level values. Confidence intervals for post hoc contrasts are unadjusted, whereas their p-values are Holm-adjusted. Category-specific t tests are complementary exploratory analyses.

## 4.3. Achievement by Response Format

To evaluate H3, scores for each response format were expressed as percentages of their form-specific maximum scores before calculating posttest-minus-pretest percentage-point gains. Mauchly's test indicated that sphericity was violated, W = 0.85, $\chi^2(2) = 7.21$, p = 0.027. A Greenhouse-Geisser correction was therefore applied, $\varepsilon = 0.87$. Percentage-point gains differed significantly across response formats, F(1.74, 78.18) = 8.07, p = 0.001, partial $\eta^2 = 0.15$. Mean multiple-choice performance increased from 52.90% to 64.35%, a gain of 11.45 percentage points. True-false performance increased from 61.41% to 77.72%, a gain of 16.30 percentage points, while short-answer performance increased from 27.60% to 56.13%, a gain of 28.53 percentage points.

Holm-adjusted contrasts showed that the short-answer gain exceeded the multiple-choice gain by 17.08 percentage points, adjusted p = 0.003, 95% CI [7.05, 27.11], and exceeded the true-false gain by 12.22 percentage points, adjusted p = 0.003, 95% CI [5.21, 19.24]. Multiple-choice and true-false gains did not differ significantly (adjusted p = 0.291). These results were consistent with H3.

The raw multiple-choice score difference was not statistically significant, t(45) = 0.19, p = 0.847, but this comparison is not the primary evidence because the maximum score differed between pretest and posttest. The normalised analysis instead indicated an 11.45-percentage-point increase. The larger short-answer gain should not be interpreted as direct evidence of higher-order learning or metacognitive development because response formats also differed in score range, guessing probability, item composition, and potentially item discrimination. Table 4 presents the raw and percentage-of-maximum scores used in the response-format analysis.

**Table 4.** Raw and Percentage-of-Maximum Performance by Response Format (N = 46)

| Response format | Pretest raw M (SD) | Pre max | Pretest % M (SD) | Posttest raw M (SD) | Post max | Posttest % M (SD) | Percentage-point gain |
|---|---|---|---|---|---|---|---|
| Multiple-choice | 6.35 (2.68) | 12 | 52.90 (22.31) | 6.43 (1.83) | 10 | 64.35 (18.34) | 11.45 |
| True-false | 7.37 (2.13) | 12 | 61.41 (17.77) | 9.33 (1.63) | 12 | 77.72 (13.61) | 16.30 |
| Short-answer | 20.98 (16.39) | 76 | 27.60 (21.57) | 43.78 (17.06) | 78 | 56.13 (21.87) | 28.53 |

Note. Percentages equal the raw score divided by the form-specific maximum multiplied by 100. Mauchly's W = .85, p = .027; Greenhouse-Geisser $\varepsilon$ = .87; F(1.74, 78.18) = 8.07, p = .001, partial $\eta^2$ = .15. Holm-adjusted contrasts: short-answer > multiple-choice, adjusted p = .003; short-answer > true-false, adjusted p = .003; multiple-choice versus true-false, adjusted p = .291. Percentage-point gains were calculated from unrounded participant-level values. Confidence intervals for post hoc contrasts are unadjusted.

## 4.4. Time-on-Task and Score-Time Ratio

To evaluate H4, this study compared test-completion time between pretest and posttest. Mean completion time increased from 55.26 minutes (SD = 30.54) to 72.13 minutes (SD = 13.71), representing a mean increase of 16.87 minutes, t(45) = 3.47, p = .001, dz = .51, 95% CI [7.07, 26.67]. This result was consistent with H4. The increase indicates that participants spent more time completing the posttest but does not establish whether the additional time reflected productive checking, persistence, uncertainty, overload, or inefficient processing.

For H5, the score-time ratio increased from 0.744 (SD = 0.457) at pretest to 0.845 (SD = 0.251) at posttest, corresponding to a mean change of 0.101. The difference was not statistically significant, t(45) = 1.44, p = .156, dz = .21, 95% CI [-0.040, 0.241]. The result was therefore not consistent with H5. The score-time ratio should be regarded

ISSN 2723-6471


as an exploratory indicator of performance relative to completion time rather than as a validated measure of learning efficiency. Table 5 summarizes changes in completion time and the participant-level score-time ratio.

**Table 5.** Within-Student Changes in Completion Time and Score-Time Ratio (N = 46)

| Variable | Pretest M (SD) | Posttest M (SD) | Mean difference | t(45) | p | dz | 95% CI |
|---|---|---|---|---|---|---|---|
| Completion time (min) | 55.26 (30.54) | 72.13 (13.71) | 16.87 | 3.47 | 0.001 | 0.51 | [7.07, 26.67] |
| Score-time ratio | 0.744 (0.457) | 0.845 (0.251) | 0.101 | 1.44 | 0.156 | 0.21 | [-0.040, 0.241] |

Note. Differences equal posttest minus pretest. The score-time ratio represents achievement score divided by completion time and is treated as an exploratory performance-rate indicator. Reported ratio means are means of participant-level ratios and cannot be reconstructed by dividing group mean scores by group mean completion times.

## 4.5. Summary of Findings

Overall achievement was significantly higher at posttest, and gains differed across task-difficulty categories and normalized response formats. Completion time also increased significantly, whereas the score-time ratio did not change significantly. The results were therefore consistent with H1, H2, H3, and H4 but not H5. These findings characterize within-student performance changes during the study period and do not establish causal intervention efficacy or direct changes in metacognitive regulation. Figure 3 summarizes the principal findings and their interpretation boundaries.

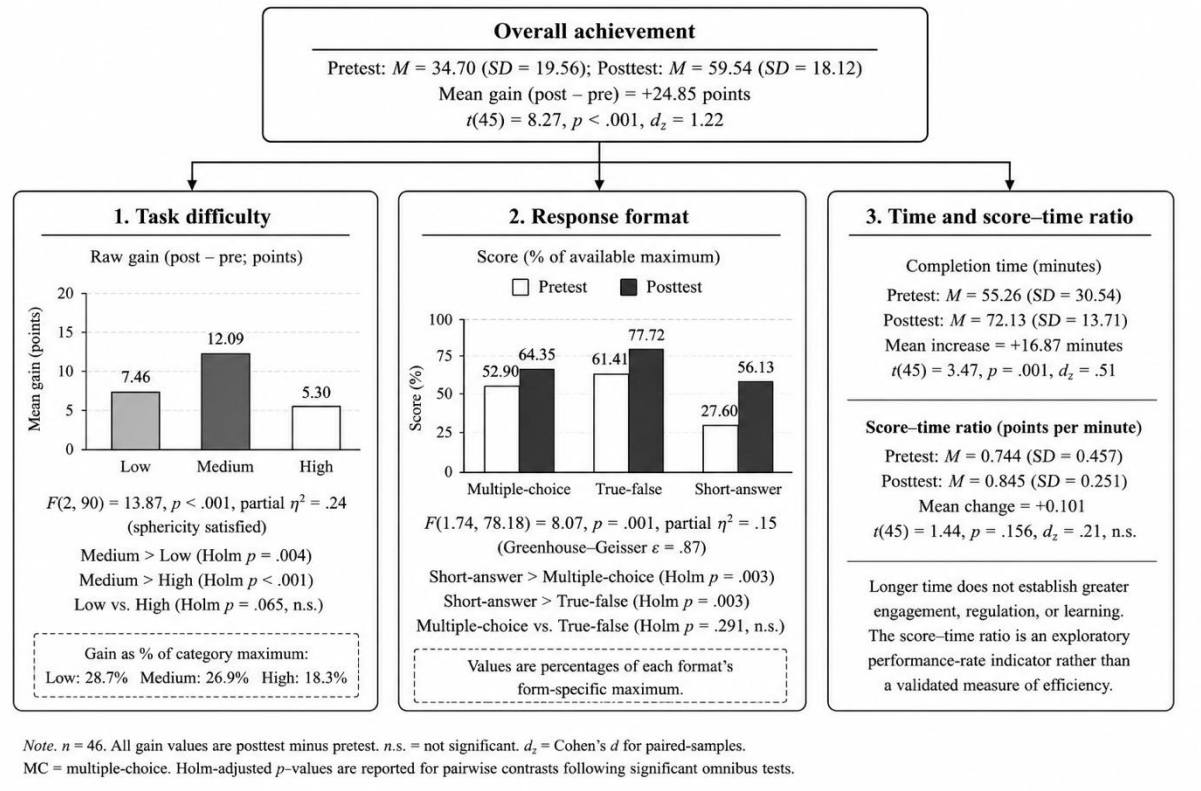



**Figure 3.** Summary of verified pretest-posttest results and interpretation boundaries.

## 5. Discussion

The findings indicate a heterogeneous pattern of within-student change rather than uniform improvement across all performance indicators. Overall achievement increased substantially, but the magnitude and interpretation of change varied across task difficulty, response format, and time-based measures. Given the one-group design and associated threats to internal validity, interpret these results as changes observed during the intervention period rather than as evidence of causal treatment efficacy.

Overall achievement increased by 24.85 points, with a large within-student effect size (dz = 1.22). This pattern aligns with a broader body of research linking metacognitive awareness, regulation, and instructional support to mathematics achievement and reasoning [6], [7], [18], [21]. Related studies have also linked metacognition with mathematical

modelling competence, reasoning, and engagement [8], [9], [10], [13]. More specifically, documented difficulties with limit concepts help explain why structured strategic support may be relevant to this demanding calculus content [4], [5]. Nevertheless, the present study did not include a comparison group, so the observed gain cannot be attributed uniquely to the intervention. The result is therefore best interpreted as substantial within-student improvement during the study period rather than as evidence that the intervention itself caused the gain.

Achievement gains also varied across task-difficulty categories. Medium-difficulty items showed the largest absolute gain, exceeding both low- and high-difficulty gains, whereas low- and high-difficulty gains did not differ significantly. One possible explanation is that medium-difficulty tasks imposed sufficient strategic demand to require planning, monitoring, and strategy selection without creating the same degree of representational and working-memory demand associated with more complex items. This interpretation is consistent with theoretical accounts emphasizing interactions among task demands, cognitive resources, metacognitive experiences, and regulatory activity [12], [25], [26]. Metacognitive problem solving may become particularly important when learners must evaluate their understanding, select among alternative strategies, and decide whether their current approach remains viable as task complexity increases [13].

In contrast, the smaller gain on high-difficulty items may reflect greater demands on working-memory resources, multi-step reasoning, and coordination across symbolic, graphical, and numerical representations. Difficulties of this kind are consistent with longstanding research showing that successful mathematical problem solving requires not only procedural knowledge but also strategic monitoring, sense making, and control of problem-solving activity [2]. These explanations remain theoretically grounded interpretations because cognitive load, mental effort, and strategy use were not directly measured in the present study. Importantly, although medium-difficulty items showed the largest absolute gain, low-difficulty items showed a slightly larger gain relative to their available score maximum. This distinction cautions against interpreting medium difficulty as an objectively optimal level of challenge.

Response-format findings likewise showed differentiated patterns of change. After normalizing scores for unequal form-specific maxima, short-answer performance showed the largest percentage-point gain and exceeded both multiple-choice and true-false gains. Multiple-choice and true-false gains did not differ significantly. One plausible explanation is that constructed-response tasks require students to generate and organize solutions, whereas selected-response formats provide response alternatives and involve greater recognition-based processing. Research comparing selected-response formats also indicates that multiple-choice and true-false questions may elicit different retrieval and comprehension processes [24]. In mathematics specifically, constructed responses may give students more opportunities to externalize reasoning, intermediate steps, and solution strategies, which are central features of mathematical problem solving [2], [27].

However, several alternative explanations warrant consideration, including guessing probability, available score range, item discrimination, test composition, and incomplete parallel-form equivalence. A simple ceiling effect is unlikely to fully explain the smaller multiple-choice gain because pretest multiple-choice performance was only approximately 53% of the available maximum. Nevertheless, the larger short-answer gain cannot be attributed uniquely to metacognitive development, deeper conceptual learning, or higher-order reasoning. Stronger conclusions about response-format effects would require psychometrically calibrated forms, including evidence of comparable item difficulty, discrimination, measurement invariance, and, where appropriate, Item Response Theory or Rasch calibration.

The time-based results further qualify the achievement findings. Students required an average of 16.87 additional minutes at posttest, while the score-time ratio increased from 0.744 to 0.845 but did not change significantly. From a metacognitive perspective, additional time could reflect more deliberate monitoring, checking, reflection, or persistence. Self-regulated learning models recognize the allocation and regulation of effort and strategy use as potentially important aspects of learning activity [16], [17], [25]. However, longer completion time can also reflect uncertainty, cognitive overload, inefficient strategy selection, or procedural slowness [25], [26]. Accordingly, completion time alone cannot determine whether students engaged in more effective regulation. Similarly, the score-time ratio should be treated as an exploratory performance-rate indicator rather than as a validated measure of learning efficiency or metacognitive regulation.

Taken together, the findings provide a descriptive and hypothesis-generating account of where pretest-posttest changes were concentrated and where no statistically significant change was detected. The overall pattern is compatible with theoretical and empirical literature linking metacognitive support with mathematics learning [18], [19], [21], but compatibility should not be confused with mechanistic evidence. The proposed regulatory pathway involving planning, monitoring, and evaluation remains theoretically plausible, yet this study did not directly demonstrate that these processes changed.

## 6. Implications

### 6.1. Theoretical Implications

The findings contribute to theory by showing that within-student performance changes varied across task difficulty, response format, and time-based indicators rather than occurring uniformly across outcomes. This pattern is broadly consistent with perspectives that conceptualize mathematical performance as emerging from interactions among task demands, strategic regulation, prior knowledge, and cognitive resources [2], [12], [25], [26]. The results therefore reinforce the importance of considering task and assessment context when examining learning outcomes associated with metacognitive instruction.

The difficulty-related findings are particularly relevant to theoretical accounts of regulation under varying cognitive demands. Planning, monitoring, and evaluation may become more consequential as learners confront tasks requiring greater strategic coordination, but increasing complexity can simultaneously impose additional demands on working memory and cognitive resources [25], [26]. The present findings are compatible with this interaction, although they do not test it directly. The larger absolute gain for medium-difficulty items should therefore be understood as a performance pattern that warrants further theoretical investigation rather than as confirmation of an optimal-difficulty mechanism.

The response-format pattern also highlights the role of assessment architecture in interpreting learning change. Different formats can impose different requirements for recognition, retrieval, solution generation, and expression of reasoning [24]. In mathematics, where competence includes strategic reasoning and monitoring in addition to obtaining a correct answer, such differences may influence what observed scores represent [2], [27]. The principal theoretical implication is therefore not that metacognitive regulation improved, but that performance change varied across task characteristics theoretically relevant to cognitive and regulatory demands.

More broadly, these findings reinforce an important distinction between metacognition as a theoretical mechanism and metacognition as an empirically demonstrated outcome. Existing research consistently associates metacognitive awareness and regulation with mathematics achievement, reasoning, and modelling performance [6], [8], [9], [10], but achievement improvement itself is not a direct measure of metacognitive development. Complementing this literature, a cross-sectional study of Ethiopian university students directly measured metacognition and reported that it partially mediated the association between emotional intelligence and academic achievement, while academic resilience moderated the relationship between emotional intelligence and metacognition [28]. Future theory testing should therefore incorporate measures that can distinguish changes in performance from changes in regulatory processes.

### 6.2. Implications for Instruction

The findings provide a rationale for further evaluating instructional approaches that embed planning prompts, progress checks, error analysis, self-questioning, and reflection within calculus problem solving. These practices align with established metacognitive instructional approaches that make mathematical reasoning and strategic regulation more explicit [18], [19], [20], [22]. Prior empirical and synthesis research has also associated metacognitive instructional support with stronger mathematics achievement [7], [21], while research on students' conceptions of limits identifies conceptual difficulties that such support may help address [4], [5].

For instructional practice, structured questioning may be particularly useful when it prompts students to clarify what a problem requires, connect relevant mathematical concepts, select a strategy, monitor intermediate steps, and reflect on the plausibility of a completed solution. These functions are consistent with metacognitive-questioning frameworks emphasizing comprehension, connection, strategy, and reflection [20]. Related research also suggests that embedding

metacognitive questions within mathematics tasks can support more deliberate engagement with problem-solving processes [22].

However, the present study does not establish that these practices are superior to conventional instruction because it included no comparison condition. Instructors should therefore interpret the findings as supporting the instructional plausibility of such practices rather than demonstrating their causal effectiveness.

When metacognitive scaffolding is used, support may be better reduced based on demonstrated student readiness than through a rigid, uniform fading schedule. Self-regulated learning theories emphasize the gradual development of learners' capacity to assume greater responsibility for planning, monitoring, and evaluating their own learning [16], [17]. Research on explicit self-regulatory strategy instruction likewise suggests that structured support can facilitate the development of increasingly independent learning behaviors [23]. In practice, instructors could therefore examine whether students can explain strategy choices, detect errors, justify intermediate steps, and evaluate completed solutions before withdrawing external prompts.

Time-on-task should not be used in isolation as an indicator of productive engagement. Additional working time may reflect careful checking, persistence, or deliberation, but it may equally reflect overload, uncertainty, or inefficient processing [25], [26]. Where feasible, instructors should interpret completion time alongside students' explanations, solution traces, reflective comments, observations, or other evidence of strategy use rather than assuming that longer working time represents more effective learning.

## 6.3. Implications for Curriculum and Assessment

At the curriculum and assessment level, the findings support using multiple response formats rather than privileging a single assessment mode. Mathematical competence includes conceptual reasoning, strategic decision-making, monitoring, justification, and solution generation in addition to answer selection [2], [27]. Constructed-response tasks can therefore complement recognition-based formats by providing students with opportunities to demonstrate intermediate reasoning and organize complete solutions.

At the same time, selected-response formats also serve legitimate assessment functions, and differences among multiple-choice, true-false, and constructed-response tasks should not be interpreted as a simple hierarchy of cognitive quality. Research indicates that even multiple-choice and true-false formats can differ in the comprehension and retrieval processes they elicit [24]. Consequently, the present results do not show that short-answer formats are intrinsically superior.

The more defensible curriculum implication is that assessment programmes should use varied formats aligned with intended learning outcomes while maintaining psychometric comparability across occasions. Where pretest-posttest comparisons are required, score maxima, item difficulty, discrimination, content coverage, guessing opportunities, and form equivalence should be considered explicitly. Format-specific differences should therefore be interpreted as potentially reflecting both learning and assessment characteristics.

## 6.4. Limitations and Future Research

Several limitations constrain interpretation of the findings. First, the one-group pretest-posttest design did not include random assignment or a comparison condition. The observed changes therefore remain vulnerable to alternative explanations, including maturation, repeated testing, concurrent instruction, instructor-related influences, routine course progression, and regression toward the mean. These threats preclude causal attribution. Future research should use randomised or well-matched comparison-group designs and include longer-term follow-up to determine whether observed gains persist and exceed changes under conventional instruction.

Second, the study involved only 46 students from one institution and identified lower-performing participants using a pragmatic cohort-relative bottom 30% rule. This threshold was not a standardized definition of low mathematical attainment, and alternative cutoffs could have produced different samples and effect estimates. No cutoff-sensitivity analysis was available. The cohort-relative definition also limits comparability with studies using standardized classifications of mathematical performance. Replication across institutions, academic programmes, and student populations is therefore needed, together with prospective evaluation of alternative screening thresholds.

Third, the study lacked direct process measures of the regulatory mechanisms proposed in its theoretical framework. The study did not directly assess planning, monitoring, evaluation, cognitive load, mental effort, or item-level response latency. This limitation matters because research conceptualizes metacognition as involving regulatory processes that cannot be inferred reliably from achievement scores alone [12], [14]. Similarly, studies associating metacognition with mathematics achievement, modelling, and reasoning typically rely on measures that explicitly operationalize metacognitive awareness or related constructs [6], [8], [9], [10]. Future studies should therefore combine achievement outcomes with validated metacognitive inventories, think-aloud protocols, observational coding, coded learning artifacts, or other direct indicators of strategy use.

Fourth, the achievement forms were not fully psychometrically calibrated. The pretest and posttest differed in response-format maxima, and evidence was unavailable for formal content-validity indices, inter-rater agreement for difficulty classification, item-discrimination statistics, cut-score validity, parallel-form equivalence, Item Response Theory or Rasch calibration, and measurement invariance. These limitations matter for interpreting the response-format findings because differences among multiple-choice, true-false, and short-answer outcomes may partly reflect assessment architecture rather than learning alone. Future studies should employ psychometrically matched forms and explicitly test measurement comparability before attributing differences to response-format-specific learning.

Fifth, the authors did not independently assess implementation fidelity. The study did not include independent observations, session-level adherence checklists, dosage records, or independent fidelity scoring. Consequently, the intended fading sequence cannot be verified empirically, and variation in implementation cannot be separated from variation in student response. Given that self-regulatory instruction commonly assumes a progressive transfer of responsibility from external support to learner control [16], [17], [23], future studies should assess whether scaffolding was delivered as intended and whether fading occurred according to demonstrated learner readiness.

Finally, completion time and the score-time ratio were relatively coarse indicators. They cannot distinguish productive deliberation from uncertainty, cognitive overload, or inefficient processing. Future research should incorporate more fine-grained temporal measures, such as item-level response latencies, together with mental-effort ratings and behavioral indicators of checking or revision. Combining such measures with direct metacognitive evidence would permit stronger differentiation among achievement change, regulatory activity, cognitive load, and learning efficiency.

## 7. Conclusion

This study documented a substantial increase in overall calculus achievement together with differentiated pretest-posttest gains across task-difficulty categories and normalized response formats. Completion time also increased, whereas the score-time ratio did not change significantly. These findings indicate that learning changes were multidimensional rather than uniform across performance indicators.

The observed pattern broadly aligns with prior research linking metacognitive awareness and instructional support to mathematics achievement, reasoning, and problem solving [6], [18], [21]. The findings are also relevant to documented conceptual challenges in learning calculus limits [4], [5]. However, consistency with prior literature does not establish that metacognitive processes caused the observed improvements.

The results therefore describe within-student change in one cohort and should not be interpreted as evidence of causal treatment efficacy or direct improvement in metacognitive regulation. Planning, monitoring, and evaluation served as theoretically proposed mechanisms but were not empirically measured. Future controlled studies should incorporate direct measures of regulatory processes, rigorous implementation-fidelity assessment, psychometrically calibrated achievement forms, and more precise measures of cognitive effort and response time to determine whether these performance changes reflect intervention-specific effects and genuine changes in metacognitive regulation.

## 8. Declarations

### 8.1. Author Contributions

Conceptualization: X.L., S.H.T.; Methodology: X.L., S.H.T.; Software: G.N., W.W.; Validation: S.H.T., X.Z.; Formal Analysis: X.L.; Investigation: G.N., W.W.; Resources: S.H.T., X.Z.; Data Curation: X.L.; Writing – Original Draft Preparation: X.L.; Writing – Review and Editing: S.H.T., W.W.; Visualization: X.Z.; All authors have read and agreed to the published version of the manuscript.

### 8.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 8.3. Funding

This study was supported by the Teaching Reform Project on Artificial Intelligence Empowering Higher Education in Hebei Province (Grant No. 2025RGZN045).

### 8.4. Institutional Review Board Statement

Formal ethical approval was not obtained for this study. Nevertheless, all participants were informed of the study's purpose and procedures, and written informed consent was obtained before participation. Participation was voluntary, and all data were anonymized and treated confidentially.

### 8.5. Informed Consent Statement

Written informed consent was obtained from all participants.

### 8.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] C. Rasmussen, K. Marrongelle, and M. C. Borba, "Research on calculus: What do we know and where do we need to go?," *ZDM Math. Educ.*, vol. 46, no. 4, pp. 507–515, Jul. 2014, doi: 10.1007/s11858-014-0615-x.

[2] A. H. Schoenfeld, "Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics," *J. Educ.*, vol. 196, no. 2, pp. 1–38, 2016, doi: 10.1177/002205741619600202.

[3] D. Tall and S. Vinner, "Concept image and concept definition in mathematics with particular reference to limits and continuity," *Educ. Stud. Math.*, vol. 12, no. 2, pp. 151–169, May 1981, doi: 10.1007/BF00305619.

[4] J. Bezuidenhout, "Limits and continuity: Some conceptions of first-year students," *Int. J. Math. Educ. Sci. Technol.*, vol. 32, no. 4, pp. 487–500, Jul. 2001, doi: 10.1080/00207390010022590.

[5] A. P. Adiredja, "Students' struggles with temporal order in the limit definition: Uncovering resources using knowledge in pieces," *Int. J. Math. Educ. Sci. Technol.*, vol. 52, no. 9, pp. 1295–1321, 2021, doi: 10.1080/0020739X.2020.1754477.

[6] U. Aydın and M. Özgeldi, "What's metacognition got to do with the relationship between test anxiety and mathematics achievement?," *Eur. J. Psychol. Educ.*, vol. 39, no. 3, pp. 2509–2529, Jul. 2024, doi: 10.1007/s10212-024-00797-7.

[7] W. Schneider and C. Artelt, "Metacognition and mathematics education," *ZDM Math. Educ.*, vol. 42, no. 2, pp. 149–161, Apr. 2010, doi: 10.1007/s11858-010-0240-2.

[8] J. R. G. Oficiar, E. D. Ibañez, and J. T. Pentang, "Metacognitive awareness as a predictor of mathematical modeling competency among preservice elementary teachers," *Int. J. Educ. Methodol.*, vol. 10, no. 2, pp. 279–292, 2024, doi: 10.12973/ijem.10.1.1079.

[9] C. C. Tak, H. Zulnaidi, and L. K. Eu, "Mediating role of metacognitive awareness between attitude and mathematics reasoning in pre-service teachers," *Eur. J. Sci. Math. Educ.*, vol. 13, no. 2, pp. 90–102, 2025, doi: 10.30935/scimath/16116.

[10] D. Varveris, V. Saltas, and V. Tsiantos, "Exploring the role of metacognition in measuring students' critical thinking and knowledge in mathematics: A comparative study of regression and neural networks," *Knowledge*, vol. 3, no. 3, pp. 333–348, Sep. 2023, doi: 10.3390/knowledge3030023.

[11] M. Boekaerts and L. Corno, "Self-regulation in the classroom: A perspective on assessment and intervention," *Appl. Psychol.*, vol. 54, no. 2, pp. 199–231, Apr. 2005, doi: 10.1111/j.1464-0597.2005.00205.x.

[12] M. V. J. Veenman, B. H. A. M. van Hout-Wolters, and P. Afflerbach, "Metacognition and learning: Conceptual and methodological considerations," *Metacognition Learn.*, vol. 1, no. 1, pp. 3–14, Apr. 2006, doi: 10.1007/s11409-006-6893-0.

[13] K. Vo, M. Sarkar, P. J. White, and E. Yuriev, "Metacognitive problem solving: Exploration of students' perspectives through the lens of multi-dimensional engagement," *Chem. Educ. Res. Pract.*, vol. 26, no. 1, pp. 141–157, 2025, doi: 10.1039/D4RP00096J.

[14] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry," *Amer. Psychol.*, vol. 34, no. 10, pp. 906–911, Oct. 1979, doi: 10.1037/0003-066X.34.10.906.

[15] G. Schraw and D. Moshman, "Metacognitive theories," *Educ. Psychol. Rev.*, vol. 7, no. 4, pp. 351–371, Dec. 1995, doi: 10.1007/BF02212307.

[16] P. R. Pintrich, "A conceptual framework for assessing motivation and self-regulated learning in college students," *Educ. Psychol. Rev.*, vol. 16, no. 4, pp. 385–407, Dec. 2004, doi: 10.1007/s10648-004-0006-x.

[17] B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educ. Psychol.*, vol. 25, no. 1, pp. 3–17, Mar. 1990, doi: 10.1207/s15326985ep2501_2.

[18] B. Kramarski and Z. R. Mevarech, "Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training," *Amer. Educ. Res. J.*, vol. 40, no. 1, pp. 281–310, 2003, doi: 10.3102/00028312040001281.

[19] Z. R. Mevarech and B. Kramarski, "IMPROVE: A multidimensional method for teaching mathematics in heterogeneous classrooms," *Amer. Educ. Res. J.*, vol. 34, no. 2, pp. 365–394, 1997, doi: 10.3102/00028312034002365.

[20] Z. Ç. Özcan and E. Erktin, "Enhancing mathematics achievement of elementary school students through homework assignments enriched with metacognitive questions," *EURASIA J. Math. Sci. Technol. Educ.*, vol. 11, no. 6, pp. 1415–1427, 2015, doi: 10.12973/eurasia.2015.1402a.

[21] J. C. Sercenia and M. S. Prudente, "Effectiveness of the metacognitive-based pedagogical intervention on mathematics achievement: A meta-analysis," *Int. J. Instruction*, vol. 16, no. 4, pp. 561–578, Oct. 2023, doi: 10.29333/iji.2023.16432a.

[22] A. Zohar and B. Peled, "The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students," *Learn. Instruction*, vol. 18, no. 4, pp. 337–353, Aug. 2008, doi: 10.1016/j.learninstruc.2007.07.001.

[23] C. Dignath, G. Buettner, and H.-P. Langfeldt, "How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes," *Educ. Res. Rev.*, vol. 3, no. 2, pp. 101–129, 2008, doi: 10.1016/j.edurev.2008.02.003.

[24] L. Hildenbrand and J. Wiley, "Supporting comprehension: The advantages of multiple-choice over true-false practice tests," *Memory Cognition*, vol. 53, no. 8, pp. 2580–2596, 2025, doi: 10.3758/s13421-025-01726-9.

[25] A. Efklides, "Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model," *Educ. Psychol.*, vol. 46, no. 1, pp. 6–25, Jan. 2011, doi: 10.1080/00461520.2011.538645.

[26] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.

[27] J. Garofalo and F. K. Lester, Jr., "Metacognition, cognitive monitoring, and mathematical performance," *J. Res. Math. Educ.*, vol. 16, no. 3, pp. 163–176, May 1985, doi: 10.5951/jresematheduc.16.3.0163.

[28] B. M. Gebresilase, Z. Z. Elka, Y. B. Biramo, M. T. Dinsa, and E. T. Taddese, "Emotional intelligence and academic achievement among Ethiopian university students: The mediating role of meta-cognition and moderating role of academic resilience," *Acta Psychol.*, vol. 269, no. 1, pp. 1–10, 2026, doi: 10.1016/j.actpsy.2026.107577.