Journal of Mathematics Instruction, Social Research and Opinion
Vol. 5, No. 3, September 2026, pp. 2327 – 2343, https://doi.org/10.58421/misro.v5i3.1316
ISSN 2962-7842

Integrating Adaptive Testing into Flipped Mathematics Classrooms: Effects on Prospective Teachers' Critical Thinking and Reasoning Quality

Arif Abdul Haqq[1], Nurma Izzati[2], Onwardono Rit Riyanto[3], Dionisio Aquino Alves[4], Siti Rahmah[5]
[1,2,5]UIN Siber Syekh Nurjati Cirebon, Indonesia
[3]Institut Prima Bangsa, Cirebon, Indonesia
[4]Departamento de Matemática, Universidade Nacional Timor Lorosa'e, Dili, Timor-Leste

Article Info

Article history:

Received 2026-03-22
Revised 2026-04-26
Accepted 2026-06-29

Keywords:

Adaptive Test
Flipped Classroom
Mathematical Critical Thinking Skills
Moodle LMS

ABSTRACT

This study addresses the limited development of students' mathematical critical thinking ability in conventional classrooms and investigates the effectiveness of a flipped classroom supported by adaptive testing. The objective of this study is to examine the effect of this instructional approach on students' mathematical critical thinking ability. An explanatory sequential mixed-methods design with a quasi-experimental pretest–posttest control group structure was employed, involving 64 students assigned to experimental and control groups. Data were collected through pre- and posttests, interviews, and classroom observations. Quantitative analysis using an independent samples t-test revealed a statistically significant difference in posttest scores, $t(62) = 4.28$, $p < .001$, with a large effect size (Cohen's $d = 1.07$). The experimental group achieved a medium normalized gain (0.47), while the control group demonstrated a low gain (0.18). Qualitative findings indicated that students in the experimental group demonstrated more elaborated argument clarification, systematic process evaluation, relational inference and deduction, and strategic completion of missing information. In contrast, control group responses were predominantly procedural and recall-oriented. These findings suggest that the integration of flipped pedagogy and adaptive testing enhances higher-order reasoning processes and students' mathematical performance, while adaptive testing is theoretically associated with more calibrated cognitive challenge and potentially more precise ability estimation.

Corresponding Author:

Arif Abdul Haqq
Tadris Matematika, UIN Siber Syekh Nurjati Cirebon, Indonesia
Email: aahaqq@uinssc.ac.id

1. INTRODUCTION

Mathematical critical thinking has become a central concern in contemporary mathematics education, particularly in preparing prospective teachers to engage with complex reasoning, justification, and problem-solving demands. However, despite its recognized importance, the development of students' mathematical critical thinking remains suboptimal, as instruction in many contexts continues to emphasize procedural fluency over higher-order reasoning. This condition reflects a critical gap between the expected competencies of prospective teachers and the instructional practices commonly implemented in mathematics classrooms.

Mathematical critical thinking encompasses several interrelated cognitive processes, including clarification of arguments, evaluation of reasoning procedures, logical inference and deduction, and strategic reconstruction of incomplete information [1], [2]. These dimensions align with constructivist learning theory, which posits that knowledge is actively constructed through meaningful engagement and reflection. Thus, instructional environments must be designed not only to deliver content but also to facilitate structured reasoning processes that support the development of higher-order thinking.

Instructional innovation has increasingly shifted toward student-centered approaches. One prominent model is the flipped classroom, which reorganizes learning activities by relocating direct instruction to the pre-class phase and utilizing in-class time for active problem solving and collaborative reasoning [3], [4]. Previous studies have demonstrated that flipped learning can improve student engagement, conceptual understanding, and academic achievement [5]. Most existing studies primarily focus on learning outcomes rather than examining how flipped instruction influences the structure of students' mathematical reasoning processes.

Alongside pedagogical reform, assessment practices have also evolved. Computerized adaptive testing offers a dynamic measurement approach in which item difficulty adjusts according to student performance, allowing more precise estimation of ability levels [6], [7]. Prior research highlights the efficiency and accuracy of adaptive testing in educational measurement. However, these studies largely treat assessment as a standalone measurement tool, without integrating it into instructional processes aimed at fostering higher-order thinking.

Despite the growing body of research on flipped classrooms and adaptive testing, a significant research gap remains in understanding how the integration of these two approaches can jointly support and explain the development of mathematical critical thinking. Existing flipped classroom studies in mathematics primarily examine achievement outcomes or engagement indicators [3], [5], yet rarely analyze how flipped structures interact with adaptive assessment mechanisms to shape reasoning processes. Similarly, research on computerized adaptive testing has largely focused on measurement precision and psychometric efficiency [6], [7], without embedding adaptive testing within a pedagogical framework designed explicitly to cultivate higher-order mathematical thinking.

Furthermore, studies of mathematical critical thinking often assess performance through static instruments, providing outcome comparisons without examining how instructional design influences reasoning development [1], [2]. This separation between pedagogy and measurement creates a conceptual gap. Instruction is treated as the driver of learning, while assessment functions merely as an evaluator rather than as a mediating mechanism within the learning process.

What remains underexplored is how an adaptive-test–supported flipped classroom might not only enhance critical thinking outcomes but also reshape the reasoning structures through which those outcomes are produced. Addressing this gap requires a design that measures performance quantitatively while explaining the processes qualitatively.

This study addresses that gap by investigating the effectiveness of a flipped classroom model supported by adaptive testing in improving mathematical critical thinking among prospective mathematics teachers. The study adopts an explanatory sequential mixed-methods design to examine both the magnitude of improvement and the reasoning mechanisms underlying observed changes [8].

Quantitatively, the study compares pretest and posttest scores between an experimental group receiving adaptive-test–supported flipped instruction and a control group receiving conventional instruction. Improvement is examined through gain and normalized gain analyses. Qualitatively, interviews and classroom observations are used to explore how students articulate arguments, evaluate processes, construct deductions, and complete missing information within the instructional context.

The following questions guide the research:

1. To what extent does the implementation of a flipped classroom supported by adaptive testing improve students' mathematical critical thinking ability?
2. How do students' reasoning processes manifest within the adaptive-test–supported flipped classroom environment?

The findings of this study are expected to contribute both theoretically and practically. Theoretically, the study offers an integrated perspective that connects instructional design and adaptive assessment within the framework of mathematical critical thinking. Practically, it provides empirical evidence and instructional insights for educators in designing learning environments that effectively promote higher-order thinking. Ultimately, this research is expected to support the development of prospective mathematics teachers who are capable of engaging in complex and reflective mathematical reasoning.

2. METHOD

2.1 Research design

This study employed an explanatory sequential mixed-methods design, in which quantitative findings were generated through a quasi-experimental pretest–posttest control group procedure and subsequently elaborated through qualitative interpretation. The sequential logic followed Creswell and Plano Clark, where statistical results serve as the analytic starting point and are later explained through qualitative evidence [8].

The quantitative strand aimed to determine whether the flipped classroom model supported by adaptive testing significantly improved students' mathematical critical thinking ability. The qualitative strand was conducted to explain differences in improvement patterns observed across students.

2.2 Instructional Model

The instructional intervention in the experimental group was grounded in the flipped classroom framework articulated by Lo et al., who emphasize relocating direct instruction to the pre-class phase and using classroom time for higher-order mathematical engagement [3]. Students accessed instructional videos and digital materials through Moodle prior to attending class. This structure aligns with the pedagogical reorganization described by Çevikbaş and Kaiser, where in-class time is optimized for active learning and collaborative reasoning [4].

During classroom sessions, students engaged in case-based problem solving, discussion, and structured reasoning tasks intended to stimulate critical mathematical thinking. The lecturer functioned as a facilitator, guiding conceptual clarification and argument development.

The intervention was conducted over eight instructional sessions (approximately four weeks), with one flipped session per meeting. Each session consisted of a pre-class phase (video-based learning and reading materials) and an in-class phase (problem solving and discussion). The flipped structure was supported by adaptive formative testing based on computerized adaptive testing principles described [6]. The adaptive testing was implemented through the Moodle quiz module using conditional branching and question bank stratification based on difficulty levels. This mechanism aimed to provide individualized challenge and more precise ability estimation, functioning simultaneously as assessment and scaffolding within the flipped learning environment. The adaptive test functioned primarily as a formative assessment during the intervention and was not directly used as the posttest measurement instrument.

2.3 Research site and participants

Participants were prospective mathematics teacher students enrolled in the Tadris Matematika Department at UIN Siber Syekh Nurjati Cirebon. Based on the most recent quantitative dataset, the study involved 64 students, divided equally into 32 students in the experimental group and 32 students in the control group. The sampling procedure followed an intact group (quasi-experimental) design, in which existing classes were assigned as experimental and control groups without random individual assignment. Both groups completed a pretest prior to the intervention and a posttest after the instructional period. Individual gain and normalized gain (N-gain) scores were calculated for each participant.

2.3 Variables of study

Variables of the Study

The independent variable in this study was the flipped classroom model supported by adaptive testing, which reorganized instruction by relocating content acquisition to the pre-class phase and integrating adaptive formative assessment during learning. The dependent variable was students' mathematical critical thinking ability, operationalized through performance on tasks measuring argument clarification, process evaluation, deductive inference, and strategic completion of missing information. Instructional group served as the comparative factor, distinguishing students exposed to the flipped adaptive model from those receiving conventional instruction, while normalized gain scores were used to capture proportional improvement across groups.

2.4 Data collection procedures

Mathematical critical thinking ability was measured using a structured written test administered as a pretest and posttest. The instrument assessed four dimensions of mathematical critical thinking as conceptualized by Evendi and Kannadass, namely clarification of arguments, evaluation of reasoning processes, inference and deduction, and strategic completion of missing information [1], [2]. Improvement was calculated using raw gain scores and normalized gain (N-gain), computed from individual pretest and posttest scores as documented in the dataset.

2.5 Data collection procedures

Mathematical critical thinking ability was measured using a structured written test administered as a pretest and posttest. The instrument assessed four dimensions of mathematical critical thinking as conceptualized by Evendi and Kannadass, namely clarification of arguments, evaluation of reasoning processes, inference and deduction, and strategic completion of missing information [1], [2].

The instrument consisted of 8 open-ended items designed to elicit higher-order reasoning. Each item was scored using an analytic rubric ranging from 0 to 4, resulting in a maximum total score of 32. Content validity was established through expert judgment involving two mathematics education specialists, while construct validity was examined through item–indicator alignment. Reliability analysis using Cronbach's alpha yielded a coefficient of 0.82, indicating high internal consistency. Improvement was calculated using raw gain scores and normalized gain (N-gain), computed from individual pretest and posttest scores as documented in the dataset.

2.6 Data analysis

Quantitative analysis followed statistical procedures recommended by the field. Descriptive statistics were first computed to examine mean pretest, posttest, gain, and N-gain values for both groups [9].

Within-group improvement was analyzed using paired-sample t-tests to determine whether pretest–posttest differences were statistically significant. Between-group differences were examined using independent-samples t-tests comparing posttest and N-gain scores between the experimental and control groups. Statistical significance was set at $\alpha = 0.05$, and effect sizes were calculated to determine the magnitude of instructional impact.

Following the explanatory sequential design, qualitative data collection was guided by quantitative results [8]. Students from the experimental group were selected using maximum variation sampling based on N-gain categories (high, moderate, and low improvement). A total of 9 students (three from each category) were interviewed.

Semi-structured interviews explored how students experienced adaptive difficulty shifts, how pre-class preparation influenced in-class reasoning, and how they constructed mathematical arguments. Each interview lasted approximately 30–45 minutes and was audio-recorded and transcribed verbatim. Classroom observations were conducted during all intervention sessions using structured observation sheets and field notes.

Data were analyzed thematically using deductive coding based on the four critical thinking indicators and inductive coding to identify emergent reasoning patterns. To ensure trustworthiness, data triangulation (interviews and observations) and peer debriefing were employed.

Integration followed an explanatory approach in which qualitative themes were used to interpret quantitative improvement patterns. Statistical evidence of differential N-gain between groups was examined alongside qualitative accounts of adaptive engagement and reasoning development. Through this integration, the study sought to explain how the flipped classroom model [3] and adaptive testing mechanism [6] jointly contributed to the development of mathematical critical thinking ability.

3. RESULTS AND DISCUSSION

This section presents the findings in accordance with the explanatory sequential mixed-methods design. The quantitative results are first reported to determine the extent to which the flipped classroom supported by adaptive testing influenced students' mathematical critical thinking ability through descriptive statistics, inferential analysis, normalized gain, and effect size estimation. These outcomes establish the magnitude and direction of instructional impact. The discussion then incorporates qualitative evidence from interviews and classroom observations to clarify how reasoning processes unfolded during implementation, thereby providing an explanatory account of the statistical patterns identified in the initial phase.

3.1 Quantitative Findings (RQ1)

To address RQ1, "To what extent does the implementation of a flipped classroom supported by adaptive testing influence students' mathematical critical thinking ability?", descriptive and inferential statistical analyses were conducted on pretest and posttest scores from both groups. In addition to mean comparison, normalized gain analysis and effect size estimation were performed to determine the magnitude of instructional impact.

Table 1. Descriptive Statistics of Mathematical Critical Thinking Scores

Statistics	**Experimental Class**		**Control Class**	
	Pretest	Posttest	Pretest	Posttest
N	32	32	32	32
Mean	11.88	14.72	12.06	10.75
Std. Deviation	1.39	3.82	2.14	3.63
Minimum	9	7	8	7
Maximum	15	20	19	20

The descriptive results indicate divergent learning trajectories. The experimental group improved from a mean of 11.88 to 14.72, yielding a mean gain of 2.84 points. In contrast, the control group decreased from 12.06 to 10.75, resulting in a negative mean gain of 1.31 points. The higher posttest maximum score in the experimental class suggests that several students reached advanced levels of performance after the intervention.

These patterns indicate that students exposed to flipped instruction supported by adaptive testing demonstrated measurable cognitive progress, whereas conventional instruction not only failed to produce comparable improvement but was also associated with a decline in mean performance. This downward shift in the control group is pedagogically important because it suggests that routine teacher-centered learning may be insufficient to maintain students' critical thinking performance when assessment tasks demand explanation, deduction, and strategic reconstruction rather than straightforward recall.

The conventional class was less consistently engaged in reflective and self-regulated reasoning practices during the intervention period. Control-group students tended to rely on recall-based responses, brief procedural execution, and formula matching, with less evidence of systematic verification.

Table 2. Independent Samples t-test and Effect Size (Posttest Scores)

Group	N	Mean	Std. Dev	t	df	Sig. (2-tailed)	Cohen's d
Experimental	32	14.72	3.82	4.28	62	0.000	1.07
Control	32	10.75	3.63				

The independent samples t-test revealed a statistically significant difference between groups, $t(62) = 4.28$, $p < .001$. The mean difference of 3.97 points indicates a substantial separation in posttest performance.

Effect size analysis yielded Cohen's $d = 1.07$, which falls within the large category. This magnitude reflects a strong practical impact, indicating that the observed difference is educationally meaningful rather than merely statistically detectable. The effect size suggests that a considerable proportion of students in the experimental group outperformed their peers in the control group at the conclusion of the intervention.

Table 3. Normalized Gain (N-Gain) Comparison

Group	N	Mean N-Gain	Category
Experimental	32	0.47	Medium
Control	32	0.18	Low

The normalized gain analysis reinforces these findings. The experimental class achieved an average N-Gain of 0.47, classified as medium improvement, whereas the control class reached 0.18, categorized as low.

To examine proportional impact, the effect size for N-Gain was also calculated. The resulting Cohen's d value of 0.98 indicates a large magnitude difference in learning growth between groups. This confirms that the flipped classroom supported by adaptive testing enhanced not only absolute performance but also proportional cognitive development relative to initial ability levels.

Across descriptive statistics, inferential testing, normalized gain classification, and effect size estimation, the results consistently indicate that the flipped classroom supported by adaptive testing exerts a strong instructional influence on mathematical critical thinking ability. The convergence of statistical significance and large effect size provides robust empirical support for the effectiveness of the intervention and establishes a quantitative foundation for the subsequent explanatory qualitative analysis.

3.2 Qualitative Findings (RQ2)

The qualitative findings examine how mathematical critical thinking was enacted during classroom interaction and individual reasoning. Rather than merely cataloguing correct responses, the analysis focuses on how students articulated arguments, evaluated processes, constructed deductions, and generated missing information. The themes below are organized according to the four indicators of mathematical critical thinking assessed in this study.

1) Clarification of Arguments and Statements

The first theme concerns students' ability to clarify arguments and articulate mathematical statements beyond surface-level recall. This indicator does not simply measure whether students know a property, but whether they can justify and elaborate on the reasoning behind it. The interviews reveal differences in how students framed explanations, especially when prompted to go beyond rule-based responses.

One experimental student explained, *"An isosceles triangle has two equal sides and two equal angles. I think the base angles are equal because the triangle is symmetrical"* (SE1). This statement shows intuitive structural reasoning. The student recognizes symmetry as a governing idea, suggesting that conceptual relationships are being perceived rather than memorized mechanically. However, the explanation remains implicit and lacks reference to formal geometric principles such as congruence or reflective symmetry. The student demonstrates emerging analytical thinking but has not yet consolidated it into a theorem-based justification.

Another student stated, *"All triangles have angles that add up to 180 degrees. If one angle is 90, the other two must add to 90. That's the rule"* (SE2). Here, the student correctly applies the angle sum principle and contextualizes it in a specific case. The reasoning is procedurally coherent, yet epistemically closed. By referring to it as "the rule," the student frames the knowledge as an accepted fact rather than derived reasoning. The explanation signals confidence in usage but limited engagement with foundational proof.

In the control group, one student remarked, *"The sum of angles is 180 because that's what we learned before"* (SC1). This response indicates reliance on prior instruction rather than reconstruction of logic. The statement reflects recall without expansion. Unlike the experimental students, there is minimal attempt to elaborate or connect ideas. Observation data confirm that in the experimental class, students were more inclined to extend explanations when prompted (OBS1), whereas control class responses tended to conclude quickly unless further questioned (OBS2).

Collectively, the data suggest that experimental students were more willing to attempt clarification, even if incomplete, whereas control students were more likely to treat mathematical statements as fixed information.

2) Evaluation of Process and Systematization

The second theme concerns how students evaluated their own procedural reasoning and organized solution steps systematically. This indicator reflects metacognitive awareness and the ability to monitor reasoning rather than merely execute formulas.

One experimental student described, *"First, I identify the base and the height. Then I multiply them. After that, I check again to make sure I didn't mix up the numbers"* (SE1).

The student explicitly frames problem-solving as a sequenced process. The act of rechecking is articulated as part of the reasoning cycle, not as an afterthought. This reveals awareness that mathematical reasoning requires validation. The explanation reflects structured thinking rather than linear execution.

Another student noted, *"Before calculating, I check if the values are correct from the question"* (SE2). This indicates that verification begins even before computation. The student demonstrates sensitivity to the accuracy of input data and alignment between the problem statement and formula. Such behavior reflects procedural discipline and internal regulation. It suggests that reasoning is treated as something to manage carefully rather than complete quickly.

In contrast, a control student stated, *"If the result looks right, I usually continue to the next problem"* (SC2). The phrase "looks right" suggests intuitive plausibility rather than structured validation. The reasoning process appears outcome-driven rather than process-aware. Observation notes confirm that verification in the control group was often reactive and prompted by lecturer feedback rather than initiated independently (OBS3). This theme indicates that experimental students more frequently articulated reasoning as a monitored process, whereas control students tended to treat procedures as tasks to be completed.

3) Inference and Deduction

The third theme focuses on deductive reasoning and the ability to manipulate mathematical relationships logically. This indicator examines whether students treat formulas as static tools or dynamic structures that can be reorganized.

One experimental student explained, *"If I know the circumference, I divide it by 2π to get the radius"* (SE1). The reasoning demonstrates algebraic transformation rather than direct substitution. The student isolates the variable conceptually, indicating understanding of equation structure. The explanation reflects awareness that formulas represent relationships that can be reorganized. This suggests relational thinking rather than memorized procedure.

Another student elaborated, *"If I know the area, I rearrange the formula. So r equals the square root of L divided by π. After I get r, I multiply by two to get the diameter"* (SE3). The reasoning unfolds in multiple inferential stages. Each step logically builds on the previous one. The student constructs intermediate conclusions before reaching the final answer. This layered reasoning reflects confidence in manipulating symbolic relationships.

In the control group, one student commented, *"I use the formula that matches the question"* (SC1). The statement emphasizes formula selection rather than transformation. The reasoning appears linear and application-focused. Observation data show fewer discussions of alternative solution pathways in the control class (OBS4). Overall, deductive reasoning in the experimental group appears more flexible and structurally articulated, whereas in the control group it remains largely substitution-based.

4) Strategy in Completing Missing Information

The final theme concerns how students responded when problems lacked direct information. This indicator reflects adaptive reasoning and the ability to generate intermediate steps independently.

One experimental student stated, *"If I only know the perimeter, I divide it by four to get the side, then I use s squared to find the area"* (SE1). The student identifies a missing variable and constructs it deliberately. The reasoning reflects planning and recognition of dependency among variables. The solution path is generated rather than immediately given. This indicates strategic reconstruction.

Another student explained, *"The height wasn't there, so I used sine to calculate the area"* (SE2). The response shows flexibility in shifting conceptual frameworks. The student moves from geometric reasoning to trigonometric reasoning when needed. Such flexibility suggests that mathematical knowledge is organized as interconnected systems rather than isolated formulas.

In the control group, hesitation was more common when required data were absent. Observation notes indicate pauses and requests for clarification when key variables were not explicitly provided (OBS2). The pattern suggests that experimental students were more likely to attempt reconstruction of missing information, whereas control students more frequently waited for complete data before proceeding.

Across all four themes, differences emerge in reasoning depth, monitoring, transformation, and adaptive strategy use. Experimental students articulate longer explanations, verify procedures more explicitly, reorganize formulas with greater flexibility, and generate intermediate information more confidently. Control students often demonstrate correct answers but shorter reasoning chains and less visible metacognitive regulation.

This section presents the qualitative findings descriptively. The following section integrates these patterns with quantitative results to explain differential learning outcomes.

3.3 Integration Findings (RQ3)

The integration phase explains how the statistical differences identified in RQ1 are supported by the reasoning patterns observed in RQ2. Quantitative results showed significant posttest differences, large effect sizes, and higher normalized gain in the experimental group. These indicators confirm strong instructional impact, yet they do not by themselves explain how that improvement occurred.

The qualitative findings provide the explanatory layer. Differences emerged in how students clarified arguments, monitored procedures, transformed formulas, and reconstructed missing information. The joint display below aligns these process-level patterns with statistical outcomes to demonstrate how shifts in reasoning depth and metacognitive regulation correspond to measured gains in mathematical critical thinking.

Table 5. Joint Display of Quantitative Results and Qualitative Explanations

Quantitative Results (RQ1)	Qualitative Evidence (RQ2)	Integrated Interpretation
Post-test Mean Difference: Experimental = 14.72; Control = 10.75 t(62) = 4.28, p < .001 Cohen's d = 1.07 (Large)	Experimental students articulated structural reasoning: *"I think the base angles are equal because the triangle is symmetrical."* (SE1) Control students relied on recall: *"The sum of angles is 180 because that's what we learned before."* (SC1) OBS1 shows an extended explanation in the experimental class; OBS2 shows brief responses in the control.	The large effect size is consistent with deeper conceptual articulation in the experimental group. Flipped learning with adaptive testing appears to shift reasoning from recall-based justification toward structural explanation, increasing conceptual clarity and argumentative elaboration.
Mean Gain: Experimental = +2.84 Control = −1.31	Experimental students described sequenced verification: *"After that, I check again to make sure I didn't mix up the numbers."* (SE1) *"Before calculating, I check if the values are correct."* (SE2) Control student: *"If the result looks right, I usually continue."* (SC2) OBS3 confirms reactive verification in the control group.	Positive gain in the experimental group aligns with explicit metacognitive monitoring. Improvement is not merely score-based but process-based. The negative gain in the control group corresponds with outcome-oriented reasoning lacking systematic verification.
Normalized Gain: Experimental = 0.47 (Medium) Control = 0.18 (Low) N-Gain Effect Size d = 0.98 (Large)	Experimental students reorganized formulas: *"I divide it by 2π to get the radius."* (SE1) *"I rearrange the formula… then multiply by two."* (SE3) Control student: *"I use the formula that matches the question."* (SC1) OBS4 shows limited alternative pathway discussion in the control.	The large N-Gain effect reflects proportional cognitive restructuring. Experimental students treat formulas as relational systems that can be manipulated. Control students apply formulas statically. Growth appears tied to structural flexibility rather than formula recognition alone.
Higher Posttest Maximum (20) in Experimental Group	Experimental students reconstructed missing variables: *"If I only know the perimeter, I divide it by four to get the side."* (SE1) *"The height wasn't there, so I used sine."* (SE2) OBS2 reports hesitation in the control group when the data is incomplete.	The emergence of higher peak performance corresponds with adaptive strategy generation. Students in the experimental group demonstrate strategic reconstruction under incomplete conditions, indicating higher-order transfer and flexible reasoning.

The joint display reveals a consistent pattern of convergence between statistical magnitude and qualitative depth. Each quantitative indicator is mirrored by a distinct shift in how reasoning was enacted in the classroom. The integration, therefore, does not merely juxtapose numbers and narratives; it shows structural alignment between performance growth and cognitive processes.

First, the large posttest difference (Cohen’s d = 1.07) corresponds with differences in argumentative clarification. Experimental students attempted structural explanations, invoking symmetry and relational properties, even when their justifications were not fully formalized. Control students, by contrast, framed statements as remembered rules. This distinction is subtle but consequential. The statistical separation in scores reflects more than accuracy; it reflects expansion in explanatory effort. Higher mean performance aligns with greater willingness to elaborate and connect ideas conceptually.

Second, the divergence in mean gain (+2.84 vs. −1.31) is coherently explained by differences in procedural monitoring. Experimental students described sequenced reasoning and explicit verification, indicating that checking and validating were internalized as part of problem solving. Control students relied on intuitive plausibility, often proceeding when answers “looked right.” The negative gain in the control group is not an anomaly but consistent with reasoning that is outcome-driven rather than process-regulated. Improvement appears tied to metacognitive discipline rather than task completion speed.

Third, the large effect size in normalized gain (d = 0.98) is mirrored in students’ treatment of formulas. Experimental students reorganized equations, isolated variables, and articulated multi-step transformations. Control students selected formulas that “matched” the question but rarely manipulated them structurally. The proportional growth captured by N-Gain reflects cognitive restructuring. When formulas are understood as relational systems rather than fixed templates, learning extends beyond substitution into flexible deduction.

Finally, the higher posttest maximum in the experimental class corresponds with adaptive reconstruction strategies. When information was incomplete, experimental students generated intermediate variables and shifted conceptual frameworks, including trigonometric reasoning when necessary. Observation data show that control students hesitated under similar conditions. Peak performance, therefore, appears linked to strategic autonomy. Students who can generate missing information demonstrate transfer, not just retention.

Across all rows, the table indicates explanatory coherence. Statistical improvement is consistently associated with deeper articulation, monitored reasoning, symbolic flexibility, and adaptive strategy use. The flipped classroom supported by adaptive testing did not merely increase scores; it altered the structure of reasoning enacted during problem solving. This alignment strengthens the causal plausibility of the intervention and provides a process-based explanation for the observed quantitative effects.

3.4 Discussion

The findings of this study indicate that the flipped classroom supported by adaptive testing generated substantial improvement in students’ mathematical critical thinking. The large posttest effect size and meaningful normalized gain are not isolated statistical artifacts;

they are consistent with qualitative evidence showing deeper clarification of arguments, structured verification, relational manipulation of formulas, and adaptive reconstruction of missing information. The convergence between measured performance and enacted reasoning suggests that the intervention influenced the structure of cognition rather than merely accelerating procedural fluency. Despite the promising findings, the relatively small sample size and the use of a single institutional context limit the generalizability of the results. The observed effects should therefore be interpreted as context-bound rather than universally representative.

These results are theoretically aligned with research demonstrating that students learn more effectively when they apply knowledge rather than recall procedures [3], [10]. Conceptual problems require learners to recognize relevant mathematical ideas independently [11], [12], to analyze and model relationships [13], and to articulate reasoning in ways that extend beyond algorithmic substitution. The experimental group's increased use of relational explanations and multi-step transformations reflects this shift. Such reasoning corresponds with the broader claim that flipped instruction reallocates classroom time toward higher-order thinking and problem solving [5], [14].

The mechanism underlying this shift appears to involve conceptual integration. Learning is strengthened when new knowledge is connected to prior beliefs and schemas [3]. The qualitative data show students deriving intermediate variables, reorganizing formulas, and shifting between geometric and trigonometric frameworks. These practices illustrate integration rather than repetition. Peer explanation and dialogic exchange, both associated with flipped learning environments, are also known to enhance understanding through argumentative clarification [15], [16]. The observed expansion in explanatory chains suggests that students increasingly treated mathematical knowledge as a system of relations. The limited duration of the intervention constrains the extent to which long-term impacts on critical thinking development can be inferred. The findings primarily reflect short-term cognitive gains rather than sustained transformation.

The adaptive testing component introduces an additional explanatory dimension. Adaptive assessment selects items matched to the learner's estimated ability [17], [18], and is theoretically associated with maximizing informational value and improving measurement precision [19], [20]. By aligning difficulty with ability, adaptive tests theoretically maintain appropriate cognitive challenge [18], [21]. The sustained engagement and extended reasoning observed in the experimental group are consistent with this principle. Furthermore, adaptive performance is evaluated based on which items are answered correctly, not merely the quantity of correct responses [18]. This emphasis on response quality parallels the study's focus on reasoning depth rather than surface accuracy.

Methodologically, the use of adaptive structures resonates with broader international trends. Large-scale assessments such as PISA have implemented multistage adaptive testing to improve precision and fairness across ability levels [22], [23]. The emphasis on accurate ability estimation and minimization of measurement error [6], [19] reflects a shift away from static total-score logic, which may obscure comparability [24]. When adaptive logic is embedded within instruction, it is theoretically expected to enhance cognitive engagement and may potentially contribute to more precise ability estimation, as suggested in prior literature.

At the same time, the literature cautions that adaptive testing may introduce affective challenges. Higher test anxiety can negatively influence performance in adaptive environments [18], [25], [26]. Although anxiety was not directly examined here, future research should consider emotional regulation as part of the critical thinking assessment. The validity of adaptive measurement depends not only on algorithmic precision but also on learners' psychological readiness [27], [28].

The formative dimension of adaptive systems also deserves attention. Structured formative assessment has been shown to promote achievement in flipped contexts[3]. Automatic formative assessment systems, particularly those offering interactive feedback, support reasoning at activity, process, and self-regulation levels [29], [30]. Such feedback shifts education from static knowledge transmission toward skill construction [31]. The qualitative evidence of verification and self-monitoring in the experimental group suggests that adaptive testing functioned not merely as evaluation but as formative scaffolding. It is also important to acknowledge the potential influence of instructor effects, as variations in teaching style, facilitation quality, and familiarity with flipped pedagogy may have contributed to the observed outcomes.

Broader scholarship on digital learning environments reinforces this interpretation. Active and self-regulated learning is associated with enhanced critical thinking development [32], [33]. Social interaction and cognitive presence are central to higher-order reasoning processes [34], yet purely descriptive analytic approaches may fail to capture deeper constructs [35], [36]. The integration of adaptive measurement with flipped pedagogy appears to overcome this limitation by linking interaction, reasoning articulation, and performance estimation within a coherent system.

Finally, the contextual validity of critical thinking assessment remains a persistent concern. Educational research has shown that unadapted instruments may not accurately reflect local students' real performance [7], [37], [38]. The present study contributes by situating adaptive testing within an instructional ecology rather than importing it as a detached measurement tool. By aligning pedagogy, formative feedback, and adaptive estimation, the intervention appears to have created conditions under which mathematical reasoning became both more explicit and more measurable. These limitations indicate that the findings should be interpreted with caution, and further research involving larger samples, multiple institutional settings, longer intervention periods, and controlled instructor variables is necessary to strengthen the robustness and transferability of the results.

In sum, the discussion suggests that the observed gains in mathematical critical thinking stem from instructional assessment coherence. Flipped pedagogy expanded opportunities for analysis, argumentation, and collaborative reasoning [39], [40]. Adaptive testing is conceptually designed to calibrate cognitive demand and is often associated in the literature with more precise measurement. Their interaction appears to have reshaped the structure of students' reasoning while rendering its growth statistically detectable. Future research should further examine affective moderation, long-term transfer, and contextual adaptation to ensure that such coherence remains both effective and equitable.

3.5 Conclusion

This study demonstrates that integrating a flipped classroom with adaptive testing supports the development of students' mathematical critical thinking, as reflected in more structured, reflective, and flexible reasoning processes. The findings indicate that the observed improvement is not only statistical but also process-based, highlighting the importance of alignment between instructional design and assessment mechanisms in fostering higher-order thinking. These findings should be interpreted with caution. The study was conducted with a relatively small and localized sample of prospective mathematics teachers within a single institutional context, and the intervention was implemented over a limited duration. These constraints restrict the generalizability of the results and limit conclusions regarding long-term impact.

This study reinforces the view that pedagogy and assessment function most effectively as an integrated system. From a practical perspective, the findings suggest that embedding adaptive assessment within student-centered learning environments may help sustain appropriate cognitive challenge and support deeper engagement with mathematical reasoning. Future research is needed to examine longitudinal effects, incorporate affective variables such as motivation and self-efficacy, and test the approach across more diverse educational contexts. By addressing these areas, subsequent studies can strengthen the empirical foundation and broaden the applicability of instructional–assessment integration in mathematics education.

REFERENCES

[1] E. Evendi, H. Retnawati, and E. Istiyono, "Assessment of mathematical critical thinking skills: A literature review," *BetaJTM*, vol. 16, no. 2, pp. 85–97, Nov. 2023, doi: 10.20414/betajtm.v16i2.592.

[2] P. Kannadass, R. Hidayat, P. S. Siregar, and A. P. Husain, "Relationship Between Computational and Critical Thinking Towards Modelling Competency Among Pre-Service Mathematics Teachers," *TEM Journal*, pp. 1370–1382, Aug. 2023, doi: 10.18421/TEM123-17.

[3] C. K. Lo, K. F. Hew, and G. Chen, "Toward a set of design principles for mathematics flipped classrooms: A synthesis of research in mathematics education," *Educational Research Review*, vol. 22, pp. 50–73, Nov. 2017, doi: 10.1016/j.edurev.2017.08.002.

[4] M. Cevikbas and G. Kaiser, "Can flipped classroom pedagogy offer promising perspectives for mathematics education on pandemic-related issues? A systematic literature review," *ZDM Mathematics Education*, vol. 55, no. 1, pp. 177–191, Feb. 2023, doi: 10.1007/s11858-022-01388-w.

[5] Kepler College, Kigali, Rwanda and C. Dusengimana, "Trends in the use of flipped classroom model and its effectiveness in higher learning education: A systematic review," *Afr Educ Res J*, vol. 11, no. 4, pp. 616–633, Nov. 2023, doi: 10.30918/AERJ.114.23.096.

[6] H.-H. Chang and Z. Ying, "A Global Information Approach to Computerized Adaptive Testing," *Applied Psychological Measurement*, vol. 20, no. 3, pp. 213–229, Sep. 1996, doi: 10.1177/014662169602000303.

[7] Y. Wang and Z. Wu, "Adapting or adopting? Critical thinking education in the East Asian cultural sphere: A systematic integrative review," *Thinking Skills and Creativity*, vol. 49, p. 101330, Sep. 2023, doi: 10.1016/j.tsc.2023.101330.

[8] J. W. Creswell and V. L. Plano Clark, *Designing and conducting mixed methods research (3rd ed.)*. SAGE Publications, 2018.

[9] Field, *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

[10] B. Rittle-Johnson and R. S. Siegler, "The relation between conceptual and procedural knowledge in learning mathematics: A review," in *The Development of Mathematical Skills*, 1st ed., London: Psychology Press, 2021, pp. 75–110. doi: 10.4324/9781315784755-6.

[11] E. Kennedy, B. Beaudrie, D. C. Ernst, and R. St. Laurent, "Inverted Pedagogy in Second Semester Calculus," *PRIMUS*, vol. 25, no. 9–10, pp. 892–906, Nov. 2015, doi: 10.1080/10511970.2015.1031301.

[12] R. Rupnow and T. Fukawa-Connelly, "How mathematicians characterize and attempt to develop understanding of concepts and definitions in proof-based courses," *Front. Educ.*, vol. 8, p. 1284666, Jan. 2024, doi: 10.3389/feduc.2023.1284666.

[13] R. Kirvan, C. R. Rakes, and R. Zamora, “Flipping an Algebra Classroom: Analyzing, Modeling, and Solving Systems of Linear Equations,” *Computers in the Schools*, vol. 32, no. 3–4, pp. 201–223, Oct. 2015, doi: 10.1080/07380569.2015.1093902.

[14] G. S. Mason, T. R. Shuman, and K. E. Cook, “Comparing the Effectiveness of an Inverted Classroom to a Traditional Classroom in an Upper-Division Engineering Course,” *IEEE Trans. Educ.*, vol. 56, no. 4, pp. 430–435, Nov. 2013, doi: 10.1109/TE.2013.2249066.

[15] V. Jungić, H. Kaur, J. Mulholland, and C. Xin, “On flipping the classroom in large first year calculus courses,” *International Journal of Mathematical Education in Science and Technology*, vol. 46, no. 4, pp. 508–520, May 2015, doi: 10.1080/0020739X.2014.990529.

[16] N. Rahmatika, M. H. B. M. Hanapi, J. Jendriadi, Y. Sariasih, M. K. W. Imami, and U. Uslan, “The Impact of Flipped Learning on Students` Critical Thinking; A Systematic Literature Review,” *IJLTER*, vol. 23, no. 10, pp. 180–206, Oct. 2024, doi: 10.26803/ijlter.23.10.9.

[17] F. Kezer, “The Effect of Item Pools of Different Strengths on the Test Results of Computerized-Adaptive Testing,” *International Journal of Assessment Tools in Education*, vol. 8, no. 1, pp. 145–155, Mar. 2021, doi: 10.21449/ijate.735155.

[18] H. Lu, Y. Hu, J. Gao, and Kinshuk, “The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing,” *Computers & Education*, vol. 100, pp. 45–55, Sep. 2016, doi: 10.1016/j.compedu.2016.04.012.

[19] D. Delgado-Gómez, J. C. Laria, and D. Ruiz-Hernández, “Computerized adaptive test and decision trees: A unifying approach,” *Expert Systems with Applications*, vol. 117, pp. 358–366, Mar. 2019, doi: 10.1016/j.eswa.2018.09.052.

[20] W. J. Van Der Linden, “Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing,” *Journal of Educational and Behavioral Statistics*, vol. 28, no. 3, pp. 249–265, Sep. 2003, doi: 10.3102/10769986028003249.

[21] M. D. Marraffino, B. L. Schroeder, N. W. Fraulini, W. L. Van Buskirk, and C. I. Johnson, “Adapting training in real time: An empirical test of adaptive difficulty schedules,” *Military Psychology*, vol. 33, no. 3, pp. 136–151, May 2021, doi: 10.1080/08995605.2021.1897451.

[22] K. Yamamoto, H. J. Shin, and L. Khorramdel, “Multistage Adaptive Testing Design in International Large-Scale Assessments,” *Educational Measurement*, vol. 37, no. 4, pp. 16–27, Dec. 2018, doi: 10.1111/emip.12226.

[23] K. Yamamoto, H. J. Shin, and L. Khorramdel, “Introduction of multistage adaptive testing design in PISA 2018,” OECD Education Working Papers 209, Nov. 2019. doi: 10.1787/b9435d4b-en.

[24] M. Ueno and P. Songmuang, “Computerized Adaptive Testing Based on Decision Tree,” in *2010 10th IEEE International Conference on Advanced Learning Technologies*, Sousse, Tunisia: IEEE, Jul. 2010, pp. 191–193. doi: 10.1109/ICALT.2010.58.

[25] T. M. Ortner, E. Weißkopf, and T. Koch, “I Will Probably Fail: Higher Ability Students’ Motivational Experiences During Adaptive Achievement Testing,” *European Journal of Psychological Assessment*, vol. 30, no. 1, pp. 48–56, May 2014, doi: 10.1027/1015-5759/a000168.

[26] T. M. Ortner and J. Caspers, “Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing,” *European Journal of Psychological Assessment*, vol. 27, no. 3, pp. 157–163, Jan. 2011, doi: 10.1027/1015-5759/a000062.

[27] S. L. Wise and C. E. DeMars, “Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions,” *Educational Assessment*, vol. 10, no. 1, pp. 1–17, Feb. 2005, doi: 10.1207/s15326977ea1001_1.

[28] V. D. Novak, D. Hass, M. S. Hossain, A. F. Sowers, and J. D. Clapp, “Effects of adaptation accuracy and magnitude in affect-aware difficulty adaptation for the multi-attribute task battery,” *International Journal of Human-Computer Studies*, vol. 183, p. 103180, Mar. 2024, doi: 10.1016/j.ijhcs.2023.103180.

[29] A. Barana, C. Fissore, M. Marchisio Conte, and M. Tassone, “Shaping an Adaptive Path on Analytic Geometry with Automatic Formative Assessment and Interactive Feedback:,” in *Proceedings of the 16th International Conference on Computer Supported Education*, Angers, France: SCITEPRESS - Science and Technology Publications, 2024, pp. 361–372. doi: 10.5220/0012627900003693.

[30] G. Corbalan, F. Paas, and H. Cuypers, “Computer-based feedback in linear algebra: Effects on transfer performance and motivation,” *Computers & Education*, vol. 55, no. 2, pp. 692–703, Sep. 2010, doi: 10.1016/j.compedu.2010.03.002.

[31] D. Carless, “From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes,” *Active Learning in Higher Education*, vol. 23, no. 2, pp. 143–153, Jul. 2022, doi: 10.1177/1469787420945845.

[32] A. R. Vargas-Murillo, I. N. M. D. L. A. Pari-Bedoya, and I. L. Ypanaque-Pereira, “Analyzing the impact of Moodle as an Interactive Tool to Improve Students’ Engagement and Motivation in the Learning

Process: A Systematic Literature Review," in *2023 9th International Conference on e-Society e-Learning and e-Technologies*, Portsmouth United Kingdom: ACM, Jun. 2023, pp. 46–53. doi: 10.1145/3613944.3613952.

[33] A. Dimas, "The Effect Of Self-Regulated Learning On Critical Thinking Skills And Learning Outcomes: Pengaruh Self Regulated Learning Terhadap Kemampuan Berpikir Kritis Dan Hasil BelajaR," *solidarity*, vol. 3, no. 1, pp. 18–32, Jul. 2023, doi: 10.35719/solidarity.v3i1.82.

[34] R. Kaliisa, B. Rienties, A. I. Mørch, and A. Kluge, "Social learning analytics in computer-supported collaborative learning environments: A systematic review of empirical studies," *Computers and Education Open*, vol. 3, p. 100073, Dec. 2022, doi: 10.1016/j.caeo.2022.100073.

[35] M. Dado and D. Bodemer, "A review of methodological applications of social network analysis in computer-supported collaborative learning," *Educational Research Review*, vol. 22, pp. 159–180, Nov. 2017, doi: 10.1016/j.edurev.2017.08.005.

[36] M. Saqr, U. Fors, M. Tedre, and J. Nouri, "How social network analysis can be used to monitor online collaborative learning and guide an informed intervention," *PLoS ONE*, vol. 13, no. 3, p. e0194777, Mar. 2018, doi: 10.1371/journal.pone.0194777.

[37] T. S. Betancourt *et al.*, "Assessing local instrument reliability and validity: a field-based example from northern Uganda," *Soc Psychiat Epidemiol*, vol. 44, no. 8, pp. 685–692, Aug. 2009, doi: 10.1007/s00127-008-0475-1.

[38] N. Fadilah and . M., "Design and Content Validity Analysis of Physics Test based on Local Wisdom for High School Students," *International Journal of Educational Research Review*, vol. 4, no. 4, pp. 602–609, Oct. 2019, doi: 10.24331/ijere.628426.

[39] T. Låg and R. G. Sæle, "Does the Flipped Classroom Improve Student Learning and Satisfaction? A Systematic Review and Meta-Analysis," *AERA Open*, vol. 5, no. 3, p. 2332858419870489, Jul. 2019, doi: 10.1177/2332858419870489.

[40] E. A. Van Vliet, J. C. Winnips, and N. Brouwer, "Flipped-Class Pedagogy Enhances Student Metacognition and Collaborative-Learning Strategies in Higher Education But Effect Does Not Persist," *LSE*, vol. 14, no. 3, p. ar26, Sep. 2015, doi: 10.1187/cbe.14-09-0141.