From ChatGPT to classroom learning: Exploring the role of teacher mediation in AI-supported education

Inês Borges
Polytechnic University of Coimbra, Coimbra, Portugal

Cristina M. R. Caridade
Polytechnic University of Coimbra, Coimbra, Portugal; inED – Center for Research and Innovation in Education, Polytechnic University of Coimbra, Coimbra, Portugal

Cláudia Sebastião
Polytechnic University of Coimbra, Coimbra, Portugal

Verónica Pereira
School of Technology and Management of Lamego, Polytechnic University of Viseu, Viseu, Portugal; inED –Center for Research and Innovation in Education, Polytechnic University of Coimbra, Coimbra, Portugal

This study investigated how undergraduate students in a first-year applied mathematics course engaged with artificial intelligence (AI) tools and how teacher mediation influenced learning outcomes. Adopting a lesson study approach, grounded in the framework of instrumental orchestration, the research followed 127 students as they tackled a new integration method. Results from the autonomous phase revealed a verification deficit, despite high AI usage habits, only 15.6% of groups produced correct or mostly correct solutions using tools like ChatGPT and only 14.2% of students reported feeling confident. Following a structured intervention focused on equipping students with foundational mathematical knowledge and verification strategies, Wilcoxon signed-rank tests indicated that group performance rose significantly to 74.2% ($Z = 4.62$, $p < .001$, $r = .82$) and individual confidence increased to 36.3%. Furthermore, a final problem posing phase showed that shifting students from consumers to evaluators enabled groups to construct problems through prompt refinement and validation. These findings suggest that while AI tools foster autonomy, they require explicit teacher scaffolding to transform blind trust into critical understanding. The study highlights implications for course design, teacher training and institutional strategies for integrating AI into higher education, demonstrating the value of lesson study for examining how emerging technologies reshape pedagogical practice.

Implications for practice or policy:

- First-year students require foundational structural knowledge before autonomous AI engagement to mitigate the verification deficit.
- Teacher mediation must evolve from simple content exposition to epistemic scaffolding, explicitly equipping students with the structural criteria required to audit and validate algorithmic outputs.
- Course designs should integrate problem posing tasks. Shifting students from consumers to evaluators compels them to create and refine prompts, validating structural understanding and fostering the internalisation of concepts.

Keywords: artificial intelligence, mathematics education, instrumental orchestration, teacher mediation, AI-assisted learning, lesson study

Introduction

The transition to higher education is often accompanied by increased academic demands and greater expectations of learner autonomy (Casanova et al., 2020). To cope with these challenges, many students turn to self-directed digital resources, including artificial intelligence (AI) tools that provide immediate

explanations and solutions to complex problems. The rapid adoption of generative AI systems such as ChatGPT is reshaping how learners engage with disciplinary knowledge across higher education (Crompton & Burke, 2023), particularly in mathematically intensive fields where procedural reasoning and validation are central.

While such tools may support independent study, emerging research highlights a critical limitation as novice learners often lack the evaluative judgement required to assess the quality and validity of AI-generated outputs (Bearman et al., 2024; Tai et al., 2018). In these situations, students may develop an illusion of explanatory depth (Rozenblit & Keil, 2002), accepting plausible but incorrect algorithmic solutions without engaging in systematic verification. This raises a key pedagogical question: Is autonomous AI-supported learning sufficient for conceptual understanding, or is structured teacher mediation necessary to transform algorithmic outputs into objects of critical inquiry?

This study examined the role of teacher mediation in AI-supported mathematics learning, focusing on how instructional orchestration supports students in validating and understanding AI-generated procedures. It addressed two research questions:

1. To what extent does autonomous AI-supported study enable students to verify complex mathematical results?
2. How does teacher mediation transform uncritical reliance on AI outputs into critical mathematical understanding?

To investigate these questions, a lesson study design employing a productive failure approach was implemented (Ponte et al., 2014), in which students first attempted to solve unfamiliar integration problems using AI tools and subsequently received structured instructional mediation.

In this study, the difficulty of critically validating AI-generated outputs was conceptualised as a verification deficit, understood not merely as an individual cognitive limitation but as an epistemic condition emerging from the interaction between probabilistic AI outputs and insufficiently developed validation schemes. This construct guided the design, measurement and interpretation of the intervention.

The following hypotheses were formulated to assess the impact of teacher mediation on learning outcomes as indicators of this verification deficit:

1. Performance
   H_{o1} (null): Teacher mediation does not produce significant differences in students' group performance compared with autonomous AI use.
   H_{11} (alternative): Teacher mediation significantly improves students' group performance compared with autonomous AI use.
2. Perceived confidence
   H_{o2} (null): Teacher mediation does not significantly change students' individual perceived confidence compared with autonomous AI use.
   H_{12} (alternative): Teacher mediation significantly increases students' individual perceived confidence compared with autonomous AI use.

By empirically examining the transition from autonomous AI use to mediated learning, this study contributes to the literature by demonstrating how teacher-mediated instrumental orchestration transforms generative AI from a solution generator into a verification instrument, thereby addressing the verification deficit in AI-supported learning contexts. The article proceeds with a review of relevant literature, followed by the methodology, results and conclusions discussing implications for AI-integrated curricula in higher education.

Literature review

This study is grounded in a socio-constructivist perspective of learning, in which knowledge is actively constructed through interaction with others and with available tools (Vygotsky, 1978). Within this framework, the teacher assumes a central mediating role, particularly when learners confront unfamiliar concepts or cognitively demanding tasks requiring conceptual validation. Instructional scaffolding (Wood et al., 1976) provides temporary support that enables learners to perform tasks beyond their current level of independent competence. This process becomes especially relevant in technology-rich learning environments, where students must interpret, question and evaluate automated outputs. In AI-supported contexts, effective mediation requires the integration of technological, pedagogical and disciplinary knowledge, consistent with the technological pedagogical content knowledge framework (Mishra & Koehler, 2006).

Recent advances in AI have significantly influenced learning practices in higher education mathematics, where AI tools are increasingly used to support automated problem-solving, adaptive learning systems and personalised feedback (Egara & Mosimege, 2024). Research on AI-enhanced learning environments has highlighted the importance of structured system design and teacher complementarity in supporting student learning (Holstein et al., 2019). The emergence of generative AI systems represents a substantial shift in this landscape. Unlike earlier rule-based educational technologies, generative AI systems such as ChatGPT produce context-sensitive responses based on large-scale language modelling (Crompton & Burke, 2023). Although these outputs often appear coherent and authoritative, they may contain inaccuracies that are not immediately evident to novice learners. Consequently, effective educational integration of generative AI requires domain-specific validation strategies that enable learners to interrogate and verify algorithmic outputs rather than merely reproduce them (Al-Ali & Miles, 2025).

The growing availability of generative AI tools exposes a cognitive vulnerability among novice learners that may be interpreted through the lens of automation bias, defined as the tendency to overrely on algorithmic outputs, particularly when systems appear authoritative (Romeo & Conti, 2026). In mathematics learning contexts, this bias may be amplified by the illusion of explanatory depth (Rozenblit & Keil, 2002), whereby individuals believe they understand a concept more deeply than they do. When interacting with generative AI systems, students may therefore accept syntactically coherent but conceptually flawed solutions without engaging in systematic verification processes. These dynamics create challenges for evaluative judgement, understood as the capacity to assess the quality and validity of one's own work and that of others (Tai et al., 2018). In contexts where generative AI provides immediate and authoritative-looking outputs, learners may rely on these tools to compensate for perceived gaps in domain knowledge while lacking the expertise to critically evaluate accuracy (Bearman et al., 2024). Together, these mechanisms clarify the epistemic conditions under which the verification deficit examined in this study may emerge.

To address this challenge, the present study drew on the framework of instrumental orchestration (Drijvers et al., 2010; Trouche, 2004), which distinguishes between a technological artefact and an instrument. An artefact refers to the tool itself, whereas an instrument emerges when the user develops effective ways of working with that tool. From this perspective, generative AI does not become a meaningful learning resource merely by being available in the classroom. Its educational value depends on how students learn to use it, particularly through strategies that support validation. In this study, teacher mediation is understood as deliberate guidance that supports this process. By modelling verification strategies, breaking down procedures into mathematically verifiable steps and encouraging students to check results through formal reasoning, the teacher helps students move beyond passive acceptance of AI-generated answers. In doing so, AI shifts from being primarily a provider of solutions to becoming a resource that supports deeper mathematical understanding. This perspective provides the conceptual foundation for the empirical analysis that follows.

Methodology and study design

Study context, participants and collaborative team

This study adopted a lesson study approach, a collaborative methodology based on cycles of joint planning, classroom observation and collective reflection (Hervas, 2021; Lewis, 2002; Ponte et al., 2014). We, the four authors, jointly designed the research lesson, instructional materials, observation protocol and grading rubric. The two of us responsible for the course acted as lead instructors, each implementing the intervention in their respective classes. In each session, a peer of our research team served as a non-participatory observer and did not take part in grading. To ensure procedural independence, the two of us who had neither taught nor observed the corresponding session completed the exercise grading independently.

The intervention took place in the first semester of the 2024-2025 academic year in Applied Mathematics I, a first-year course in a business management degree at Coimbra Business School, Polytechnic University of Coimbra. A total of 127 students across four classes participated, organised into 32 collaborative work groups. Each class completed two lesson study sessions (eight sessions overall). We conducted this study as part of a pedagogical improvement cycle. In accordance with institutional guidelines for classroom-based practitioner inquiry, the intervention constituted a pedagogical activity aimed at enhancing learning outcomes and therefore did not require formal ethical review.

Student participation was voluntary and framed within the course's continuous assessment structure as an optional group assignment. We informed students in advance that anonymised data from their work and survey responses could be used for research and pedagogical evaluation purposes, explicitly assuring them that non-participation would not affect their academic standing. To ensure no personal identifiers were retained in the research database, our research team fully anonymised all collected data after the grading process and prior to analysis.

The study focused on integration of rational functions by partial fractions, a topic in the integration unit where students typically experience difficulties. At the time of the intervention, students had covered only elementary integrals and integration by parts.

The AI-related objectives of the lesson study were to:

- characterise students' habits, trust and perceptions regarding AI use in mathematics learning
- identify limitations during autonomous AI-supported exploration of unfamiliar procedures, particularly verification strategies
- evaluate the impact of teacher mediation on procedural and conceptual mastery by comparing group performance before (AI-supported) and after (post-mediation, without consultation) the intervention
- analyse changes in perceived confidence between the autonomous and the post-mediation phases, particularly regarding solving exercises correctly and validating results.

The lesson study comprised two in-person sessions (2 hr 30 min and 2 hr), following the sequence presented in Figure 1. In line with the research questions and hypotheses outlined in the Introduction, we focused the analysis on Exercises 1–4, the mediation phase, Surveys 1 and 2 and the final exercise.

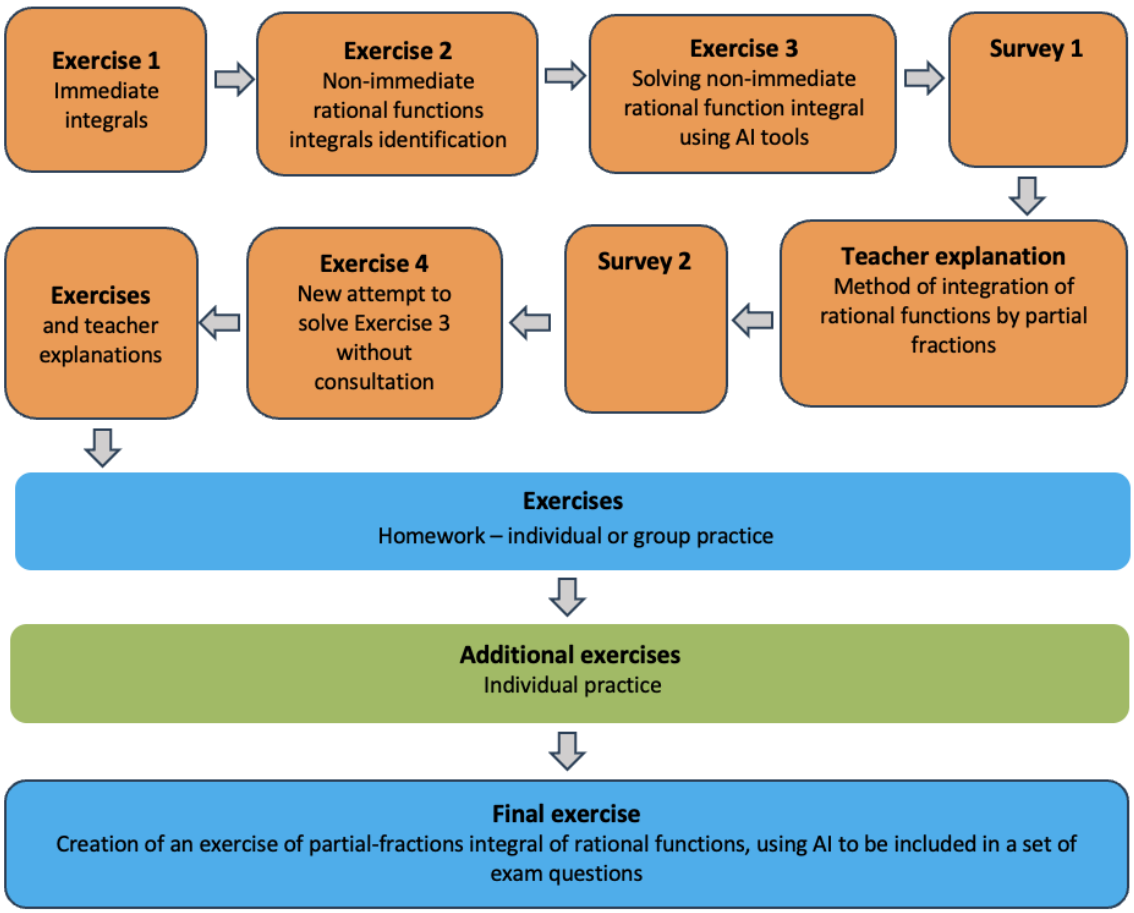

Figure 1. Sequence of activities in the lesson study: Session 1 in orange, Session 2 in green and asynchronous work in blue.

Implementation of the lesson study

During the first lesson study session, students worked in groups of four (32 groups) and completed a structured worksheet sequence designed to progressively introduce integration of rational functions by partial fractions. The initial tasks revisited prerequisite knowledge and prompted students to identify non-elementary rational integrals (Figures 2 and 3).

1. Consider the following rational functions $\frac{1}{x^2}$, $-\frac{1}{x}$, $\frac{2}{x+1}$ and $\frac{-x}{x^2+1}$.

   Do the given functions have immediate integrals? Calculate them.

Figure 2. Exercise 1 from the worksheet of Session 1

2. Now consider the rational function $\frac{x^4+1}{x^3+x^2}$.

   Does this function have an immediate integral? Can you solve it only using standard integral formulas?

Figure 3. Exercise 2 from the worksheet of Session 1

In Exercise 3 (Figure 4), students entered an unguided exploration phase. Using their own devices (smartphones, laptops or tablets), they consulted AI tools to investigate how to compute the integral and worked collaboratively to interpret the method, then determined the antiderivative. A few platforms were provided only as examples, and students could use any tools they already knew. The teacher offered no clarification or validation, so the activity functioned as an autonomous instrumental genesis process. Each group then listed the tools consulted.

3. Use AI tools (e.g., ChatGPT, Wolfram Alpha, Symbolab, Microsoft Math Solver, Photomath, Socratic...) to research how to calculate the $\int \left(\frac{x^4+1}{x^3+x^2} \right)$.

   Analyze in group the method used and then solve the integral based on the research. Indicate the AI tools used.

Figure 4. Exercise 3 from the worksheet of Session 1

Students subsequently completed Survey 1 individually, capturing AI use habits and perceptions, as well as baseline confidence following the autonomous task.

The teacher then delivered a structured mediation sequence to provide the theoretical foundations of the method while explicitly training verification strategies, covering the main stages of polynomial division (when applicable), denominator factorisation, root identification and classification, decomposition into partial fractions and integration by framing the procedure as a toolkit for auditing AI-generated outputs. To operationalise this audit, the teacher modelled validation routines combining mathematical reasoning and digital literacy, such as differentiating the obtained primitive, checking the consistency of linear systems used to determine coefficients, breaking complex AI-supported queries into smaller verifiable steps and cross-referencing generative outputs with deterministic computational platforms. This mediation aimed to shift students' reliance on AI from uncritical acceptance towards structured verification.

Students then completed Survey 2 individually on Moodle to assess post-intervention learning and confidence. The session moved to an internalisation phase in which groups recomputed the same integral from Exercise 3 (Figure 5) without external consultation, enabling comparison between AI-assisted and post-mediation performance.

4. Do you now have a clearer understanding of the method for solving the integral of the rational function $\frac{x^4+1}{x^3+x^2}$?

   Solve again the exercise 3, now considering the explanation given by the teacher (without consultation).

Figure 5. Exercise 4 from the worksheet of Session 1

During the remaining time, students solved additional non-elementary rational integrals for practice, and two further exercises were set for independent or group-based homework.

The second lesson study session focused on consolidating the procedure and included a diagnostic collaborative task in which groups designed and solved an original non-elementary rational-function integral using AI tools (Figure 6).

Final exercise

Each group must formulate an exercise involving the calculation of a non-immediate integral of a rational function, similar to the exercises solved during the Lesson Study, and in accordance with the following conditions:

- The rational function must be generated using an AI tool, with the constraint that the denominator of the fraction must have at least two real roots, one of which must have a multiplicity greater than 1;
- The submission must include both the problem statement and it resolution, and must be made within one week.

After validation, the proposed exercises and their respective solutions will be made available as worksheet.

A selection of these exercises will be included in the course assessment tests.

Figure 6. Final exercise from the worksheet of Session 2

Groups had 1 week to submit the problem statement and full solution via Moodle. Once validated, submissions were compiled into a shared practice worksheet to support preparation for upcoming assessments. This activity aimed to promote active involvement in assessment and encourage critical engagement with AI tools.

Data analysis strategies

The data collected in this study were analysed across three complementary dimensions:

- quantitative and qualitative analysis of students' exercise solutions (group level, n = 32)
- descriptive and inferential statistical analysis of survey responses (individual level, n = 123)
- qualitative analysis of classroom observations conducted by observing teachers.

Together, these dimensions supported evaluation of students' learning outcomes, perceptions and engagement throughout the lesson study. Data were pre-processed in Microsoft Excel and analysed in IBM SPSS Statistics; graphs were produced in both tools to support visual interpretation.

Analysis of exercise solutions

Exercises produced across the two lesson study sessions (Exercises 1–4 and the final exercise) were analysed with distinct objectives. Exercises 1–4 were graded at the group level (n = 32) on a 0–100 points scale using an analytic rubric designed to assess procedural accuracy, conceptual understanding and clarity of mathematical reasoning in the integration of rational functions. Analytical emphasis was placed on Exercises 3 and 4, evaluated using the rubric in Table 1, as these scores provided the main basis for comparing group performance before and after teacher mediation. Grading was completed by a member of our team who was neither instructor nor observer in the corresponding session and was cross-validated by a second independent peer to support inter-rater reliability and procedural independence.

Table 1
Rubric for evaluating group solutions to Exercises 3 and 4

Criterion	Description	Indicators of achievement	Weight (%)
Problem identification	Recognition of the type of integral and appropriate solution strategy	Correct identification of immediate vs. non-immediate rational integrals	15
Preliminary procedures	Execution of preparatory algebraic steps	Correct polynomial division when applicable and appropriate algebraic manipulation	15
Denominator analysis	Mathematical analysis required for decomposition	Correct denominator factorisation and the identification and classification of roots (real or complex, multiplicity)	20
Partial fraction decomposition	Construction of decomposition model	Correct setup of partial fractions and determination of coefficients	25
Integration procedure	Execution of integration steps	Correct integration of decomposed terms and coherent progression of calculations	15
Mathematical reasoning and presentation	Clarity and justification of reasoning	Logical justification, coherent explanation, and clear presentation of results	10
Total			100

The scoring system used continuous values to capture subtle differences in solution quality and support statistical analysis. In addition to the quantitative score, group responses were classified into three categories for qualitative interpretation – correct or mostly correct (≥70 points), minor errors (40, 70) and significant errors (<40) – corresponding to relevant conceptual or structural issues. This categorisation supported interpretation of changes in group performance, particularly when comparing Exercise 3 (AI-supported) with Exercise 4 (post-mediation, without external consultation), and enabled cross-analysis with survey data.

The final exercise was analysed with a different purpose: to verify whether each group's proposed statement met the defined criteria (a rational function whose denominator has at least two real roots, one with multiplicity greater than 1) and whether the submitted solution was mathematically correct.

Analysis of the survey's responses

Survey data were analysed using descriptive statistics (frequencies, percentages, means and medians) and presented graphically to support interpretation. From 127 enrolled students, 123 valid responses were obtained as four were excluded due to absence or incomplete submission.

Survey 1 was administered after the AI-assisted task (Exercise 3) and aimed to:

- profile students by age, gender, academic year and mathematics background
- identify habits of AI use in mathematics study, including platforms used and purposes
- assess perceptions of the effectiveness of AI-generated responses
- examine students' understanding of the solution and confidence in solving a similar problem, thereby capturing the verification deficit during the autonomous phase
- assess views on the usefulness of AI in learning contexts.

Item formats included multiple-choice questions, 5-point Likert scales (e.g., in perceived usefulness and confidence) and multiple-response items on purposes of AI use.

Survey 2 was administered after teacher mediation and focused on perceived changes in understanding and confidence when solving an equivalent non-elementary rational integral. It consisted of two items: a

5-category question measuring perceived improvement in understanding and the same 5-point confidence scale used in Survey 1, enabling paired comparison of individual confidence before and after mediation.

Research hypotheses

To guide the inferential analysis, the hypotheses formulated previously examined the effect of teacher mediation on two dimensions:

- students' performance in solving non-immediate rational integrals (group level)
- students' perceived confidence in solving an identical exercise (individual level).

These dimensions were treated as complementary expressions of the verification deficit, capturing demonstrated validation capacity (group performance) and perceived ability to critically evaluate and solve similar mathematical tasks independently (individual confidence).

Two measurement points were compared:

- autonomous resolution using AI tools (Exercise 3 and Survey 1)
- post-mediation resolution without external assistance (Exercise 4 and Survey 2).

The hypotheses, variables, units of analysis and objectives are summarised in Table 2.

Table 2
Research hypotheses, variables, units of analysis and objectives

Hypothesis	Description	Variables involved	Units of analysis	Objective
H_{01} (null – performance)	Teacher mediation does not produce significant differences in group performance compared to autonomous AI use.	Performance (Exercise 3 vs Exercise 4)	Group (n = 32)	Test whether group performance differs between autonomous AI use and post-mediation work.
H_{11} (alternative – performance)	Teacher mediation significantly improves group performance compared to autonomous AI use.			
H_{02} (null – confidence)	Teacher mediation does not significantly change individual perceived confidence compared to autonomous AI use.	Confidence level (Survey 1 vs Survey 2)	Individual (n = 123)	Test whether individual perceived confidence differs between autonomous AI use and post-mediation work.
H_{12} (alternative – confidence)	Teacher mediation significantly increases individual perceived confidence compared to autonomous AI use.			

Given the paired structure of the data and the non-normal distribution of the variables, hypotheses were tested using the Wilcoxon signed-rank test. Effect sizes (r) were calculated to estimate the magnitude of the mediation effect.

Lesson study observations

In line with lesson study methodology, direct classroom observations were conducted in each session by a designated team member who was not involved in instruction. Observations followed a structured protocol aimed at identifying behavioural and discursive indicators of the verification deficit during the autonomous AI-supported phase, as well as shifts towards explicit validation strategies and greater

epistemic autonomy after teacher mediation. The protocol operationalised the study's key constructs by linking interaction patterns with instrumental orchestration and changes in group performance (Table 3).

Table 3
Structured observation protocol used during lesson study sessions

Dimension	Analytical focus	Operational indicators	Relation to study constructs
Collaborative interaction	Group participation and dialogue patterns	Distribution of participation, negotiation of solution steps, peer explanation, conflict resolution	Collective reasoning processes
AI engagement patterns	Forms of interaction with AI tools during autonomous phase	Direct copying of outputs, acceptance without discussion, lack of validation, prompt refinement, cross-platform comparison	Verification deficit and instrumental genesis
Verification behaviours	Evidence of mathematical validation strategies	Backward differentiation, algebraic checking, plausibility discussions, identification of AI inconsistencies	Reduction of verification deficit
Autonomy transition	Shift from AI-supported to independent reasoning after mediation	Decreased AI reliance, increased internal reasoning, spontaneous validation attempts	Effect of teacher mediation
Engagement and regulation	Cognitive and motivational indicators	Persistence, questioning behaviour, expressions of uncertainty/confidence, task-focused discussion	Confidence development

To enhance trustworthiness and reduce observer bias, we analysed field notes thematically. Coding combined protocol-guided categories with patterns emerging across sessions, focusing on recurring behaviours such as hesitation, validation attempts and reliance on AI-generated outputs. We then triangulated observational data with the groups' written resolutions to examine convergence between interaction patterns and group performance outcomes.

Results and observations

This section reports quantitative results from exercise performance (group level, n = 32) and survey responses (individual level, n = 123), alongside qualitative classroom observations gathered across the eight lesson study sessions (two per class). Together, these data sources describe the shift from AI-assisted exploration to post-mediation validation practices.

Baseline characterisation: Student profile and AI usage habits

Survey 1 produced 123 valid responses. Participants had a median age of 18 (range: 17–43); 56% identified as female and 44% as male. Most students were in the first year of their undergraduate degree (76%), while 24% were retaking the course. All reported a secondary school mathematics background, and 74% had completed Mathematics A (a stronger preparation for the topics addressed in this lesson study).

AI use for mathematics study was common since about 90% of students reported having used AI tools to explore mathematics topics (Figure 7).

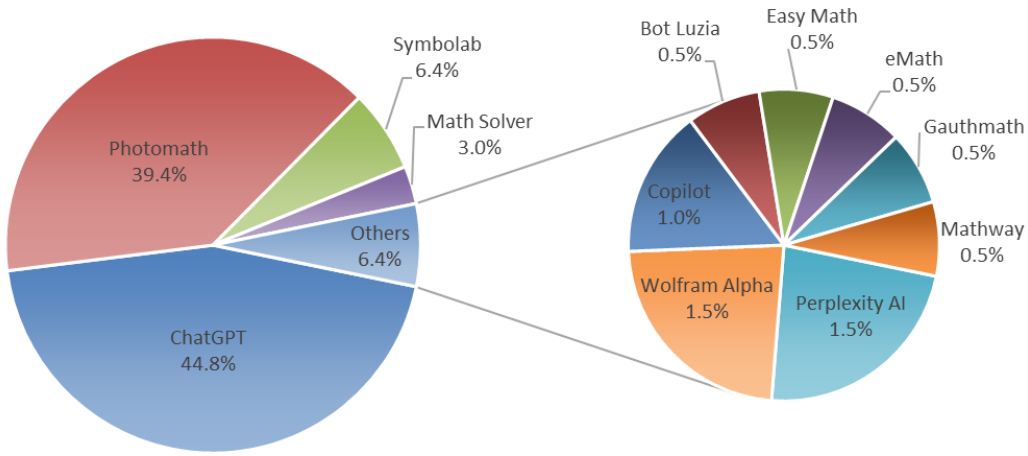

Figure 7. Pie chart of the AI tools most often cited by students when exploring mathematics topics (*n* = 123)

Students' reported purposes for these searches are summarised in Figure 8.

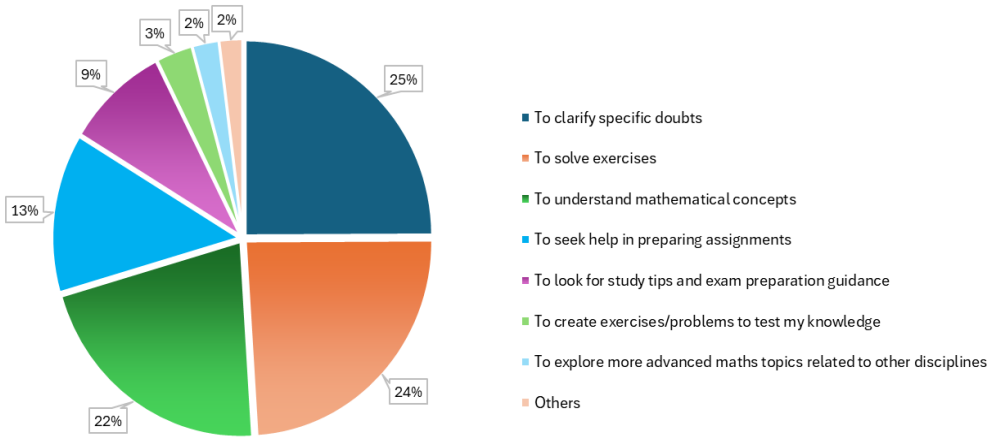

Figure 8. Pie chart of the reasons for using AI tools in mathematics learning (*n* = 123)

However, only 40.9% reported being satisfied or very satisfied with the results obtained (Figure 9).

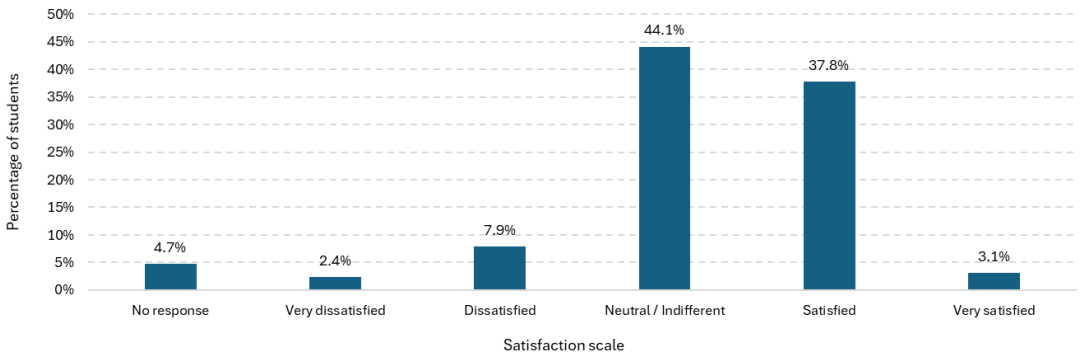

Figure 9. Bar chart of the level of satisfaction when searching mathematics topics using AI tools (*n* = 123)

Taken together, Figures 8 and 9 show that students used AI not only to clarify doubts but also to solve exercises and develop conceptual understanding. This reliance is pedagogically significant because AI explanations can appear convincing even when mathematically incorrect, particularly when students lack a solid theoretical foundation for verification. The combination of high usage and only moderate satisfaction suggests that AI use does not necessarily translate into stable confidence in the correctness of results and may be consistent with a verification deficit, understood as engagement with AI-generated explanations without well-developed validation strategies. This reinforces the relevance of examining how structured teacher mediation supports explicit validation practices.

Performance and observations during the lesson study

Exercise 1 aimed to review the concept of a rational function and to apply the standard integration rules previously taught and practised in earlier classes. Approximately 85% of the students correctly identified the type of function involved and applied the standard formulas to solve the given integrals. These results indicate that most groups were able to mobilise prior knowledge appropriately, providing a stable foundation for the subsequent tasks involving non-immediate rational integrals.

In Exercise 2, students were presented with a rational function whose integral was non-immediate. The objective was to prompt groups to recognise that the standard formulas were no longer applicable, thereby introducing the need for a different solution strategy. Most groups (94%) successfully made this distinction and attempted to solve the integral using existing techniques, such as polynomial division or rewriting the function to apply integration by parts. This preparatory phase clarified the limits of previously acquired procedures and created the conditions for the structured teacher mediation that followed.

Exercise 3, which marked the first moment of autonomous interaction with AI tools within the lesson study, revealed a diversity of approaches and success levels. Specifically, 15.6% of groups presented correct or mostly correct solutions, 40.6% exhibited minor errors, and 43.8% showed significant errors. These results indicate that a substantial proportion of groups experienced procedural or conceptual difficulties when relying exclusively on AI tools to solve a non-immediate rational integral.

The main issues identified during grading included a lack of clear justifications, omission of intermediate steps and difficulty distinguishing between essential and non-essential (or incorrect) information returned by the platforms. In several cases, groups reproduced computational steps suggested by AI systems without fully articulating the underlying mathematical reasoning.

This pattern may reflect a superficial level of procedural appropriation, in which AI-generated outputs are accepted as valid without systematic validation. Such behaviour is consistent with the verification deficit conceptualised in this study, understood as a gap between access to algorithmic solutions and the capacity to independently audit their correctness. Rather than demonstrating stable procedural mastery, many solutions suggested reliance on the tool's output without explicit evidence of structured verification.

Figure 10 presents the distribution of results for Exercise 3, grouped into the three performance categories defined in the Analysis of Exercise Solutions subsection (within the Data Analysis Strategies subsection) of the Methodology section.

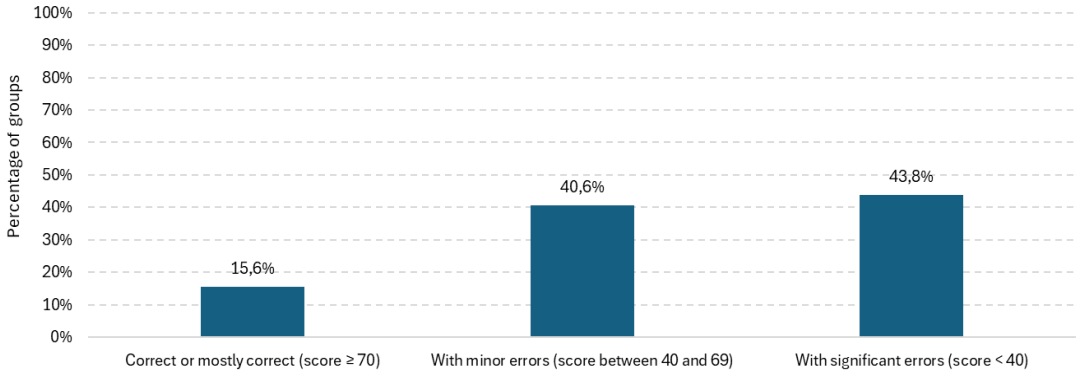

Figure 10. Bar chart of the distribution of results in Exercise 3 (*n* = 32)

In Exercise 3, the AI tools used by the various groups included ChatGPT, Photomath, Symbolab, Wolfram Alpha and Microsoft Math Solver.

During the completion of Survey 1, students were asked to assess their experience using AI tools in terms of their understanding of the solution to Exercise 3. Only 27.5% reported being satisfied or very satisfied with their understanding. These results are presented in Figure 11.

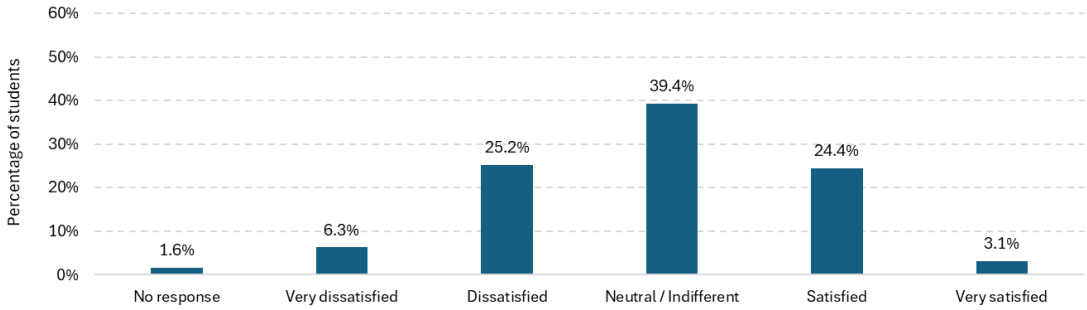

Figure 11. Bar chart of the level of satisfaction with AI-based searches in solving Exercise 3 (*n* = 123)

Regarding students' perceived confidence in correctly solving an exercise similar to Exercise 3 after using AI tools, only 14.2% reported feeling confident or very confident in their ability to do so (Figure 12).

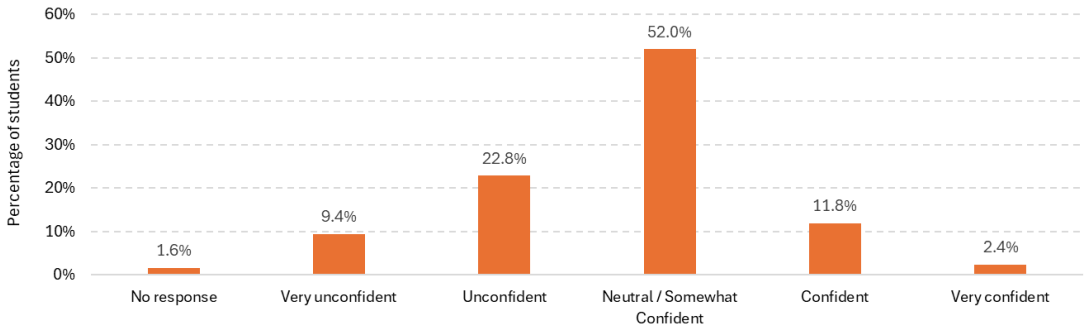

Figure 12. Bar chart of the confidence level in solving an identical exercise after conducting AI-based searches (*n* = 123)

These findings suggest that, although AI tools were widely used during the autonomous phase, relatively few students reported high levels of perceived epistemic control over the resulting solution.

In the same survey, 20% of students indicated that they did not find it meaningful to use AI tools in the context of learning mathematics in the classroom. While the survey did not directly measure the reasons for this perception, it may reflect uncertainty regarding the reliability or interpretability of AI-generated outputs.

Following the teacher mediation phase, students completed Survey 2 (n = 123 valid responses), with results presented in Figures 13 and 14.

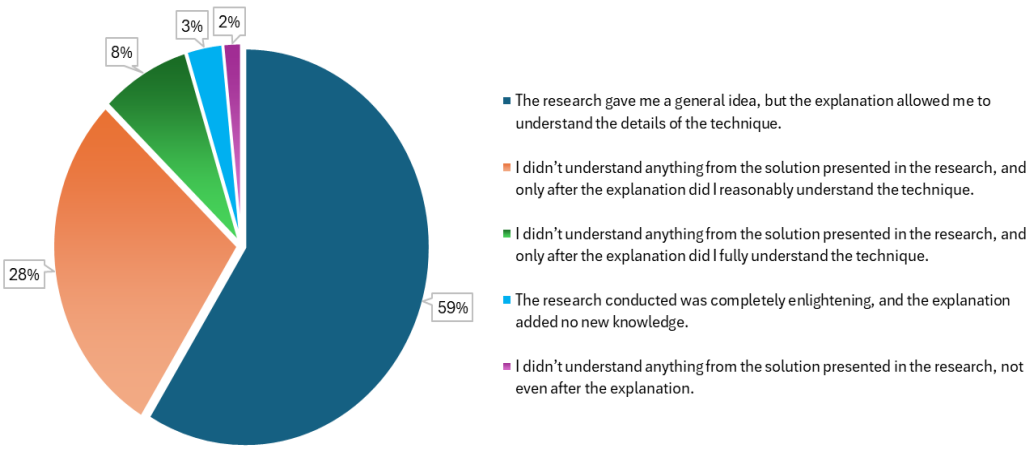

Figure 13. Pie chart of the improvement in the level of understanding of the method after teacher's mediation (n = 123)

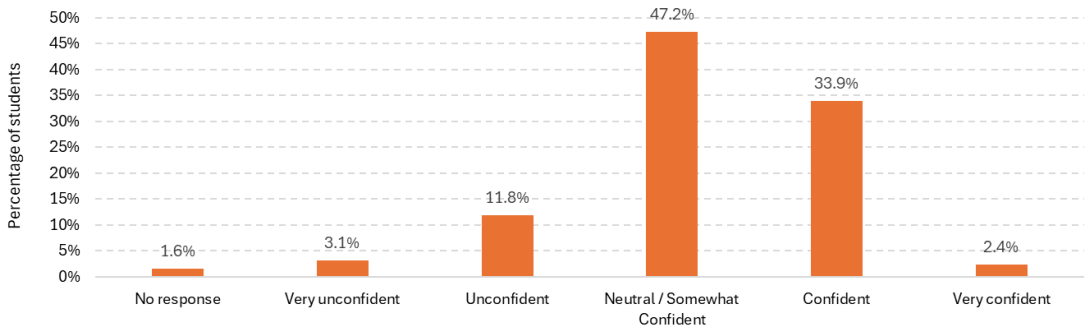

Figure 14. Bar chart of the confidence level in correctly solving an identical exercise after teacher's intervention (n = 123)

The results highlight the following:

- 95% of students indicated that the teacher's intervention contributed, to varying degrees, to an improved understanding of the method under study.
- 36.3% of students reported feeling confident or very confident in solving an exercise equivalent to Exercise 3 after mediation, compared with 14.2% in Survey 1 following autonomous AI use.

This increase in reported confidence suggests a positive shift in students' perceived ability to approach and validate similar problems following structured mediation. However, as confidence represents a subjective measure, these findings are interpreted alongside group performance data to provide a more comprehensive picture of the learning process.

In Exercise 4, students were asked to solve the same integration problem presented in Exercise 3, this time without any external support or consultation. This comparison was designed to examine whether

performance differed following the structured teacher mediation and whether the increase in reported confidence observed in Survey 2 was accompanied by measurable changes in group performance.

Figure 15 presents a ~~histogram~~ bar chart comparing the results from Exercises 3 and 4, grouped into the three performance categories defined in the Analysis of Exercise Solutions subsection (within the Data Analysis Strategies subsection) of the Methodology section.

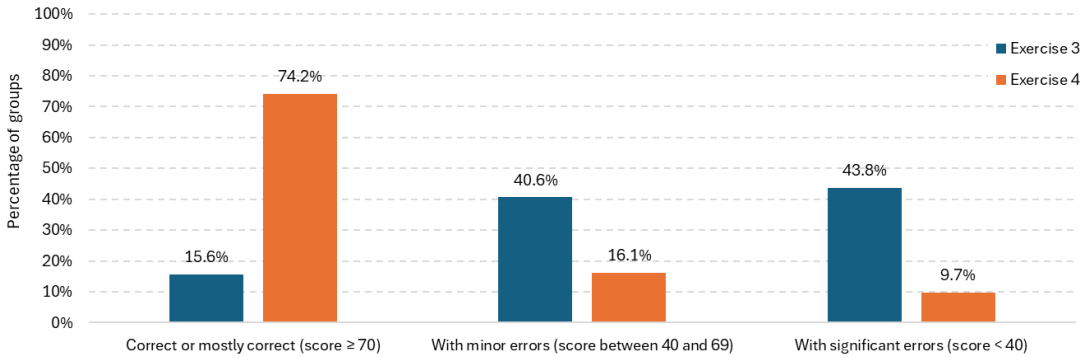

Figure 15. Comparative bar chart of the results for Exercise 3 (blue) and Exercise 4 (orange) (n = 32)

The proportion of correct or mostly correct responses at the group level rose substantially from 15.6% in Exercise 3 to 74.2% in Exercise 4. When extending this analysis to include groups with only minor errors, the proportion demonstrating at least partial procedural accuracy increased from 56.2% to 90.3%. This descriptive shift indicates a marked improvement in the quality and completeness of group solutions following mediation.

In the second class, the final exercise was introduced as a concluding task within the lesson study cycle. Each group was asked to create a non-immediate rational integral problem and provide its complete solution, using AI tools in accordance with previously defined criteria.

Of the 32 groups, 30 submitted both the problem statement and the corresponding solution. Among these:

- 24 groups fully met the required criteria and presented complete and coherent solutions that satisfied the structural conditions established in the methodology
- most groups explicitly identified the AI tools used during problem construction, indicating transparency in their process
- six submissions contained significant issues, either due to failure to meet the required criteria for problem design or because of incomplete or incorrect solutions.

These outcomes suggest that, by the end of the intervention, a substantial proportion of groups were able to use AI tools productively within defined mathematical constraints. However, a detailed analysis of the specific strategies employed in this problem-posing task falls beyond the scope of the present study and will be addressed in subsequent research.

Table 4 summarises the main indicators obtained from the different measurement instruments used throughout the study.

Table 4
Summary of the study's key indicators

Aspect analysed	Indicator and result	*n*
Prior AI experience	Students with prior experience using AI in mathematics: 90%	123
General satisfaction	Satisfaction with AI tools in mathematical searches: 40.9%	123
Initial performance	Groups with correct/mostly correct solutions (Exercise 3): 15.6%	32
Initial confidence	Students feeling confident after initial AI interaction: 14.2%	123
Impact of mediation	Students reporting improved understanding after mediation: 95%	123
Final performance	Groups with correct/mostly correct solutions (Exercise 4): 74.2%	32
Final confidence	Students feeling confident after mediation: 36.3%	123
Instrumental proficiency	Final exercises meeting all required criteria: 75%	32

Based on the analytical framework outlined in the Methodology section, changes in perceived confidence were examined by comparing students' responses after the autonomous use of AI (Survey 1) and following the teacher mediation phase (Survey 2). As confidence is an individual-level perception, the analysis included the full sample of paired responses (n = 123).

Similarly, performance in solving a non-immediate rational integral was assessed by comparing Exercise 3 (autonomous AI use) with Exercise 4 (post-mediation, without external tools). Following the lesson study methodology, this assessment was conducted at the group level (n = 32) using the pre-defined rubric (Table 1) to generate a continuous score from 0 to 100.

Before applying the comparative tests, the normality of the paired differences between the two moments was tested. The Shapiro-Wilk test indicated that both the performance differences and the confidence level differences did not follow a normal distribution (p < .001 in both cases). Therefore, the Wilcoxon signed-rank test was used, which is appropriate for this type of non-parametric, paired samples. Summaries of the respective tests are presented in Tables 5 and 6 below.

Table 5
Descriptive statistics and Wilcoxon signed-rank test for individual perceived confidence (n = 123, H_{12})

Moment	Median	IQR	Wilcoxon *W*	*Z*	*p* value	*r*
Confidence level (Survey 1 – autonomous AI use)	3	1	1950.5	5.06	<.001	.46
Confidence level (Survey 2 – after teacher mediation)	3	1				

The comparison revealed a statistically significant difference between the two moments. Although the median score remained at 3.00 (IQR = 1.00, representing the interquartile range) in both stages, the Wilcoxon signed-rank test supported the rejection of the null hypothesis (H_{02}), indicating a significant difference (W = 1950.5, Z = 5.06, p < .001).

Furthermore, the calculated effect size (r = 0.46) represents a medium-to-large impact, supporting the alternative hypothesis (H_{12}). The analysis of the ranks showed that:

- 56 students (45.5%) increased their confidence level (positive ranks)
- 12 students (9.8%) showed a decrease (negative ranks)
- 55 students (44.7%) maintained the same level of confidence (ties).

This positive shift, despite the stable median, suggests that a substantial proportion of students reported higher confidence levels after mediation.

Table 6
Descriptive statistics and Wilcoxon signed-rank test for group mathematical performance (n = 32, H_{11})

Moment	Median	IQR	Wilcoxon *W*	*Z*	*p* value	*r*
Performance in exercise 3 (autonomous AI use)	40.50	36.0	457.0	4.62	<.001	.82
Performance in exercise 4 (after teacher mediation)	90.00	34.9				

The analysis of students' group performance between the two revealed a statistically significant difference between the autonomous AI phase and the post-mediation phase. The Wilcoxon signed-rank test showed a clear improvement in the scores for Exercise 4, completed after the mediation phase, compared to Exercise 3, which was solved autonomously using AI tools (W = 457.0, Z = 4.62, p < .001, r =.82).

Furthermore, the median score more than doubled, rising from 40.50 (IQR = 36.00) to 90.00 (IQR = 34.90), supporting the rejection of the null hypothesis (H_{01}). The rank analysis further confirmed this upward trend:

- A total of 29 groups (90.6%) improved their performance (positive ranks).
- Only one group (3.1%) performed worse (negative ranks).
- Two groups (6.3%) maintained the same score (ties).

The very large effect size (r = .82) indicates a substantial magnitude of difference between the two performance measures. Given the paired design of the study, this result suggests that group performance was strongly associated with the mediation phase. While causal inferences should be interpreted cautiously in the absence of a control group, the consistency of the improvement across groups supports the interpretation that structured mediation played an important role in enhancing procedural accuracy and validation practices.

The box plots presented in Figure 16 illustrate the distributional differences between the two measurement points for both individual perceived confidence (n = 123, top) and group performance (n = 32, bottom). Regarding confidence, although the median remained stable at 3, the distribution shows an upward shift in responses, with fewer observations concentrated at the lower end of the scale. This pattern is consistent with the Wilcoxon results indicating a statistically significant difference between the two moments. In terms of group performance, the contrast between the two stages is visually pronounced. The median increased from 40.50 to 90.00, and the distribution of scores in the post-mediation phase appears more concentrated in the higher range of the scale. The absence of outliers in the second stage suggests greater homogeneity in group performance following mediation.

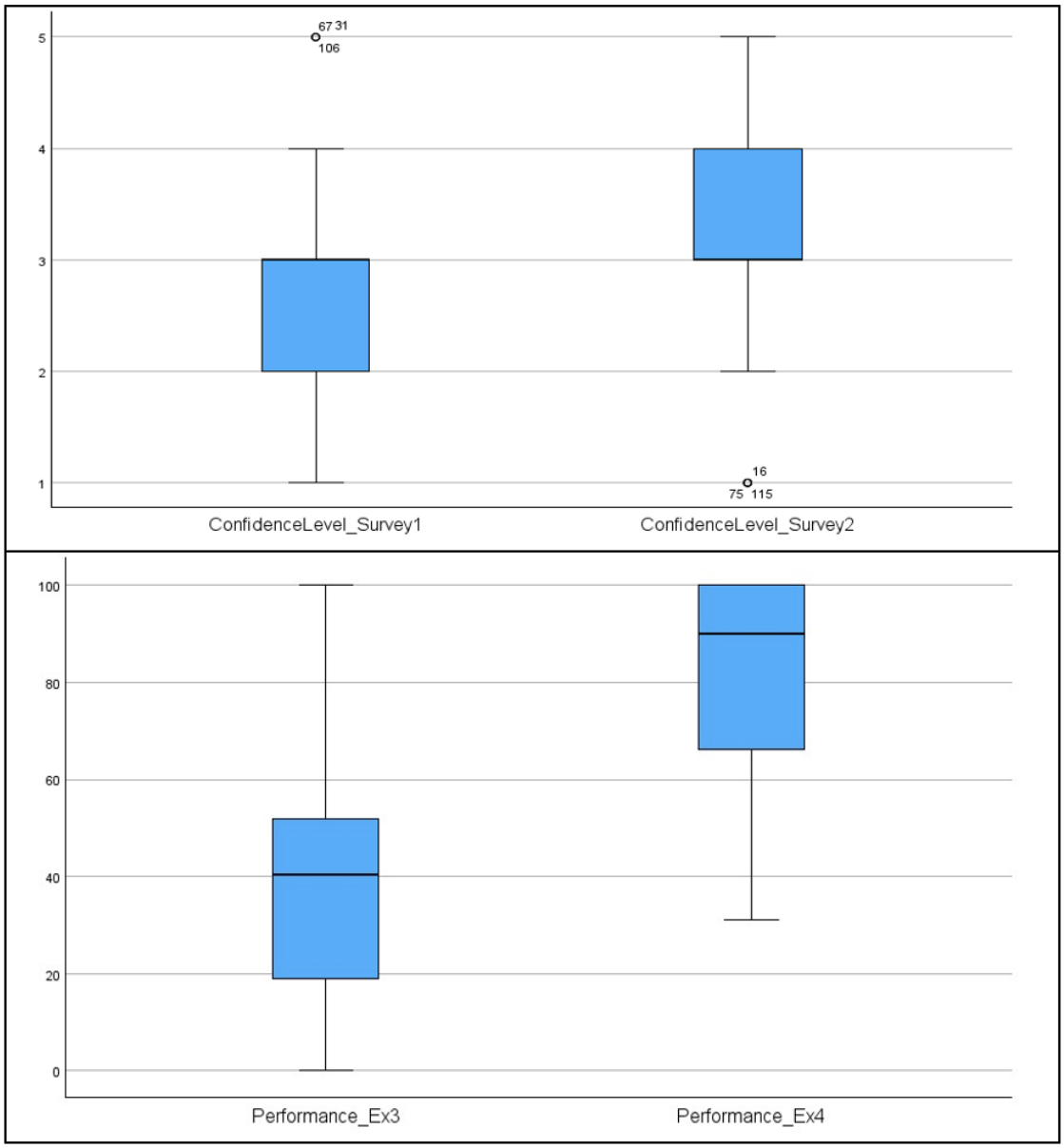

Figure 16. Boxplots comparing individual perceived confidence (*n* = 123, top) and group performance (*n* = 32, bottom) before and after teacher mediation

Qualitative observations

Field observations conducted by those of us acting as observing teachers indicated high levels of engagement throughout the sessions. From the beginning of the first session, students showed motivation and curiosity regarding the adopted methodology, particularly the opportunity to use AI tools in the study of mathematics. The introduction of a new topic, combined with the use of digital resources and collaborative work, contributed to a dynamic and participatory learning environment. Most groups engaged actively in both the autonomous exploration phase using AI and the subsequent systematisation phase led by the teacher.

The observational data, analysed through different dimensions (Table 3), revealed distinct behavioural patterns across these phases. Our observations during Exercise 3 provided multiple indicators consistent with the verification deficit identified in the study's framework. During the autonomous AI-supported phase, students frequently accepted AI-generated outputs with limited critical interrogation. We noted a

recurrent pattern of rapid scrolling, where students bypassed explanatory text to focus primarily on the final mathematical expression. Our field notes captured this tendency clearly, with one group remarking, “This result looks really complicated ... but the steps make sense. Just write it down” (Group 7).

In another instance, a group attempted to cross-reference results using two different platforms. Faced with conflicting outputs, they expressed difficulty determining which was mathematically valid, stating, “We tried two different apps and got different results. Both look logical, so we don’t know which one to pick” (Group 23). Such episodes illustrate the challenges students experienced when required to evaluate AI-generated solutions without established validation criteria. These interaction patterns were consistent with the distribution of errors observed in Exercise 3.

We observed anoticeable shift in discourse and behaviour during Exercise 4. Following teacher mediation, students demonstrated greater use of explicit validation strategies. We recorded instances of backward differentiation, peer checking of algebraic steps, and discussion of decomposition structure. For example, Group 12 stated, “Wait, if we differentiate this result, we don't get the original function. Probably we must have failed the partial fraction decomposition. Let’s check”. This change in dialogue aligns with the substantial increase in group performance (Mdn = 90.00) observed in the post-mediation phase.

The proposal of the final exercise was received with visible engagement. Our field notes indicate that students perceived the challenge of designing a problem to test AI as a meaningful shift in their role. Although task execution occurred asynchronously, our analysis of submitted work suggests that many groups moved beyond simple AI-generated output. Several submissions included manual justifications complementing AI-assisted solutions, indicating increased attention to validation and procedural transparency.

Discussion

The results of this study provide insights into AI integration in higher education mathematics. Survey data showed that approximately 90% of students had previously used AI tools to explore mathematical topics, confirming their widespread use in independent study. However, high adoption did not translate into procedural accuracy or validated understanding when AI was used without structured guidance.

During the autonomous phase, only 15.6% of groups submitted correct or mostly correct solutions to Exercise 3 (*Mdn* = 40.50). The contrast between moderate satisfaction (40.9%) and low procedural accuracy suggests misalignment between perceived and demonstrated understanding. This pattern is consistent with automation bias (Romeo & Conti, 2026) and metacognitive overestimation, including the illusion of explanatory depth (Rozenblit & Keil, 2002), which together can make plausible AI explanations appear correct without sufficient conceptual grounding.

Our qualitative observations support this interpretation. Students often prioritised final expressions over intermediate reasoning and struggled to adjudicate between conflicting AI outputs. These behaviours align with the verification deficit conceptualised in this study, reflecting limited capacity to validate algorithmic solutions independently.

Following teacher mediation, group performance improved substantially. In Exercise 4, completed without AI assistance, 74.2% of groups submitted correct or mostly correct solutions (*Mdn* = 90.00), with a statistically significant increase (r = .82). Although causal inference is limited by the quasi-experimental design, the magnitude and consistency of improvement suggest that structured mediation supported the development of explicit validation strategies. Our observational data also indicate increased use of backward differentiation, algebraic checking and peer validation, suggesting a shift towards applying mathematical criteria rather than relying on algorithmic authority.

A nuanced pattern emerges for perceived confidence. Although the median confidence level remained stable at 3, responses became more concentrated, with fewer extreme values. The stable median

combined with significant rank redistribution suggests recalibration of self-assessment rather than inflation. This may reflect more realistic appraisal of task complexity following exposure to validation strategies, consistent with utilisation scheme refinement in instrumental genesis (Drijvers et al., 2010).

The final exercise further illustrates this progression. Most submitted tasks met the structural criteria, and several groups complemented AI-generated outputs with manual justification. While deeper analysis of problem-posing warrants future work, these findings suggest more deliberate engagement with AI tools.

Overall, this study contributes to emerging research on AI in higher education by illustrating how teacher-mediated instrumental orchestration may reposition generative AI from a passive solution provider to an object of critical validation. Rather than eliminating AI from mathematical learning environments, the findings indicate that structured integration is important for mitigating the verification deficit observed during autonomous AI use. Theoretically, these findings extend the instrumental orchestration framework by showing how generative AI changes the way students internalise mathematical tools. Since AI often presents uncertain results with an air of authority, the teacher's role must move beyond traditional support. It becomes a form of epistemic regulation, guiding students to question, check and ultimately validate what the machine produces, ensuring that technology serves as a catalyst for critical thinking rather than a substitute for it.

Recommendations

The findings suggest several pedagogical implications for integrating generative AI in higher education mathematics.

First, autonomous use of AI tools, particularly for unfamiliar procedures, may not be sufficient to ensure validated understanding. AI should therefore complement, not replace, formal instruction within structured learning design. Teacher mediation is central in modelling validation strategies, including backward differentiation, algebraic verification, and critical comparison of outputs across platforms.

Second, AI literacy should extend beyond technical familiarity. Students need explicit training in epistemic evaluation: how to interrogate AI-generated results, identify inconsistencies and distinguish syntactic plausibility from conceptual correctness. Instrumental orchestration (Drijvers et al., 2010) offers a productive framework for this purpose.

Third, the problem-posing phase suggests that role inversion strategies, in which students design tasks to test AI outputs, may promote deeper conceptual engagement. Future curricular designs could explore structured AI-assisted problem creation to reinforce verification practices and mathematical reasoning.

Finally, confidence patterns indicate that interventions may support calibration of self-perception, not only performance gains. Learning environments that make validation criteria explicit may help students develop more realistic and stable epistemic judgements.

Study limitations

We acknowledge several limitations in this study.

First, we conducted the research within a single institutional context and focused on one specific mathematical topic (integration of rational functions by partial fractions). Although the results provide insight into AI-supported learning dynamics, generalisation to other mathematical domains or educational contexts should be approached cautiously.

Second, the quasi-experimental design did not include a control group. Improvements observed between the autonomous and mediated phases cannot be attributed exclusively to teacher mediation without considering potential maturation or task repetition effects. Although the magnitude of change was substantial, causal inference remains limited.

Third, we assessed performance at the group level, whereas we measured confidence individually. While this reflects the lesson study design, it introduces different units of analysis that should be considered when interpreting the results.

Fourth, our measurement of the verification deficit relied on behavioural indicators, performance outcomes, and self-reported confidence rather than on a dedicated psychometric instrument. Future research could benefit from developing validated scales specifically designed to assess AI-related epistemic validation capacity.

Finally, our analysis of the final exercise was limited to structural and procedural criteria. A more detailed qualitative examination of students' AI prompting strategies and reasoning processes would provide deeper insight into the development of instrumental genesis in AI-supported environments.

Author contributions

Inês Borges: Conceptualisation, Methodology, Project administration, Resources, Data curation, Formal analysis, Investigation, Writing – original draft, Visualisation, Writing – review and editing; **Cristina Caridade:** Conceptualisation, Data curation, Investigation, Writing – original draft, Writing – review and editing; **Cláudia Sebastião:** Conceptualisation, Data curation, Investigation, Writing – review and editing; **Verónica Pereira:** Conceptualisation, Data curation, Investigation, Writing – review and editing.

Acknowledgements

The first author acknowledges the Coimbra Business School for granting a release from teaching duties for scientific research during the first semester of the 2025-2026 academic year, which supported the preparation of this manuscript. Furthermore, the second and fourth authors would like to thank the Center for Research and Innovation in Education for their support throughout this research.

References

Al-Ali, S., & Miles, R. (2025). Upskilling teachers to use generative artificial intelligence: The TPTP approach for sustainable teacher support and development. *Australasian Journal of Educational Technology*, *41*(1), 88–106. https://doi.org/10.14742/ajet.9652

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Casanova, J. R., Araújo, A. M., & Almeida, L. S. (2020). Dificuldades na adaptação académica dos estudantes do 1.º ano do Ensino Superior [Difficulties in academic adaptation for first-year university students]. *Revista E-Psi*, *9*(1), 165–181.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*, Article 22. https://doi.org/10.1186/s41239-023-00392-8

Drijvers, P., Doorman, M., Boon, P., Reed, H., & Gravemeijer, K. (2010). The teacher and the tool: Instrumental orchestrations in the technology-rich mathematics classroom. *Educational Studies in Mathematics*, *75*(2), 213–234. https://doi.org/10.1007/s10649-010-9254-5

Egara, F. O., & Mosimege, M. (2024). Exploring the integration of artificial intelligence-based ChatGPT into mathematics instruction: Perceptions, challenges, and implications for educators. *Education Sciences*, *14*(7), Article 742. https://doi.org/10.3390/educsci14070742

Hervas, G. (2021). Lesson Study as a faculty development initiative in higher education: A systematic review. *AERA Open*, *7*. https://doi.org/10.1177/2332858420982564

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Lecture notes in computer science: Vol. 11625. Artificial intelligence in education* (pp. 157–171). Springer. https://doi.org/10.1007/978-3-030-23204-7_14

Lewis, C. (2002). *Lesson study: A handbook of teacher-led instructional change*. Research for Better Schools.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Ponte, J. P., Quaresma, M., Baptista, M., & Mata-Pereira, J. (2014). Os estudos de aula como processo colaborativo e reflexivo de desenvolvimento profissional [Lesson studies as a collaborative and reflective process of professional development]. In J. Sousa & I. Cevallos (Eds.), *A formação, os saberes e os desafios do professor que ensina Matemática* (pp. 61–82). Editora CRV.

Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & Society*, *41*(1), 259–278. https://doi.org/10.1007/s00146-025-02422-7

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Trouche, L. (2004). Managing the complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for Mathematical Learning*, *9*(3), 281–307. https://doi.org/10.1007/s10758-004-3468-5

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

---

Corresponding author: Inês Borges, borges@iscac.pt

Please cite as: Borges, I., Caridade, C. M. R., Sebastião, C., & Pereira, V. (2026). From ChatGPT to classroom learning: Exploring the role of teacher mediation in AI-supported education. *Australasian Journal of Educational Technology*, **(*), 1–22. https://doi.org/10.14742/ajet.11056