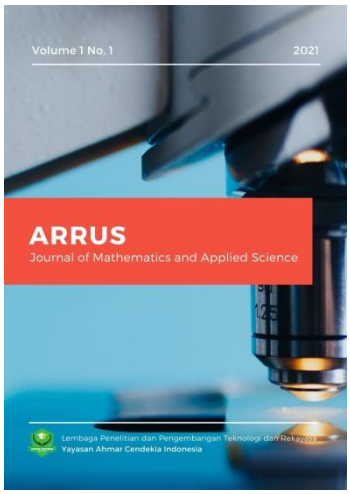
https://doi.org/10.35877/mathscience5482

OPEN ACCESS

ISSN : 2776-7922 (Print) / 2807-3037 (Online)



REVIEW ARTICLE

Co-Thinker or Cognitive Crutch? A Scoping Review of Human-AI Collaboration in University Mathematics

Irajuana Haidar[1*] & Dian Ulfa Sari[1]

[1]Mathematics Education Study Program, Universitas Sembilanbelas November Kolaka, Kolaka, Southeast Sulawesi, Indonesia.

*Corresponding author: Irajuana Haidar, Universitas Sembilanbelas November Kolaka, Indonesia

E-mail: irajuanahaidar93@gmail.com

Abstract: University students of mathematics now have a general-purpose assistant that produces worked solutions on request, yet it remains unclear whether this help strengthens or replaces self-regulated learning. This scoping review maps what empirical studies report about generative AI use, self-regulation, metacognition, verification, and cognitive offloading in university mathematics. A structured web search conducted in November 2026 identified twelve studies, published between 2024 and 2026, that met predefined criteria. Each study was charted for setting, design, sample, and findings, and one reviewer coded the role that AI played in the reported learning process. Five studies described AI as a co-thinker whose output students checked and rewrote, three as an adaptive scaffold, three as a cognitive crutch, and one showed both patterns in two students. Evidence for critical engagement came mostly from small studies in which verification was required by the course or the researchers. Evidence for offloading came from observational data in which AI use was not observed directly. No study linked observed AI use to unassisted performance in the same students, and forethought was the least observed phase of the self-regulation cycle. The review offers three testable hypotheses, one for each phase. The findings are exploratory and cannot support causal claims.

Keywords: Artificial Intelligence, Cognitive Offloading, Higher Education, Human-AI Collaboration, Self-Regulated Learning

1. Introduction

A student who reaches a stuck point in calculus or linear algebra can now paste the problem into a chatbot and receive a complete, fluent solution within seconds. Fluency is not correctness. Green (2024) read ChatGPT's answer to a typical calculus problem closely and showed how a response can display competence while exposing limits that a novice would struggle to see. Classroom reports sound more hopeful. Torres-Peña et al. (2024) described a calculus course in which several AI tools were used to deepen the understanding of derivatives and rates of change, and they observed that students became more accurate in derivative calculations. Whether such gains reflect learning or borrowed performance is the question this review takes up.

The concern is wider than mathematics. Reviews of large language models in education praise their capacity to personalise content and warn that learners need the competencies to judge their output (Kasneci et al., 2023), and students themselves report both benefits and

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

worries (Chan & Hu, 2023). A systematic review of 14 studies linked over-reliance on AI dialogue systems to weaker decision-making, critical thinking, and analytical reasoning (Zhai et al., 2024). A survey of 666 participants found that frequent AI use went with lower critical thinking, with cognitive offloading as the mediating factor (Gerlich, 2025). A survey of 319 knowledge workers found that higher confidence in generative AI went with less critical thinking (Lee et al., 2025). Experimental work in mathematics sharpens the point. In a field experiment with nearly 1,000 high school students, access to GPT-4 raised practice performance. Students who then lost access scored 17% below those who never had it, and a tutor version built with learning safeguards largely removed the harm (Bastani et al., 2025). Pardos and Bhandari (2024) found that ChatGPT-generated hints produced learning gains in algebra and statistics comparable to human tutor hints, although the hints failed quality checks on 32% of problems. Walkington (2025) reviewed the field across problem solving, tutoring and feedback, task adaptation, and teacher planning, and drew attention to its ethical and logistical issues.

Self-regulated learning describes how learners plan, monitor, and evaluate their own learning. Zimmerman (2002) organised these processes into three phases, forethought, performance, and self-reflection, that repeat from task to task, and a comparison of six SRL models concluded that such models form a coherent framework for research and teaching (Panadero, 2017). Generative AI can touch every phase. It can help a student frame a goal, answer a question during work, or offer a comparison solution afterwards. It can also remove the effortful step that a phase is meant to exercise. An et al. (2026) reviewed 39 studies of student use of generative AI in higher education. They found that students using it as a self-regulatory supporter reported the lowest ratio of challenges to successes, and that the few challenges traced to weak integrated self-regulation skills and prompt strategy. That review covered higher education broadly and did not isolate mathematics.

Mathematics adds a feature that other disciplines show to a lesser degree. A claim is defensible or it is not, and a student can in principle check it. Verification therefore becomes a learning activity in its own right, and unverified acceptance becomes a visible failure. The evidence on what students actually do is scattered across proof-based courses, calculus, linear algebra, and teacher education, and it uses different designs and outcome measures. In the search reported below, no review was found that brings these studies together around the self-regulation cycle.

This article reports a scoping review of empirical studies on generative AI in university mathematics that offer evidence on self-regulation, metacognition, verification, reliance, or offloading. Three questions guide it. RQ1: In what settings and with what designs has the topic been studied? RQ2: What do the studies report about how students plan, monitor, verify, and rely on AI output? RQ3: How strong is the evidence, and where are the gaps? The review also turns the pattern of evidence into hypotheses that later studies can test. It does not test them. It is a scoping review and not a systematic review of a bibliographic database, and the method section states what follows from that choice.

2. Literature Review

2.1. Self-regulation when the tutor is a language model

Zimmerman (2002) described self-regulation as a cycle. In forethought, the learner analyses the task and sets goals. In the performance phase, the learner applies strategies and monitors progress. In self-reflection, the learner judges the result and adjusts before the next attempt. A chatbot can support or displace each phase. A student can ask it to break a problem into steps before starting, which supports forethought. A student can ask it to explain a stuck step, which supports performance. A student can paste a finished solution and ask whether it is right, which supports self-reflection. The same three requests can also collapse into one request for the answer, and then no phase is exercised. The research question is therefore not whether students use the tool but which of these uses they choose and what follows from each.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

2.2. Evidence on AI and self-regulation outside mathematics

Cognitive offloading is the use of physical action to reduce the cognitive demand of a task. Risko and Gilbert (2016) noted that people offload when they believe it is more efficient than relying on their own ability, and that such beliefs do not always match reality. Fan et al. (2025) gave the idea an educational name. In a randomised experiment, learners supported by ChatGPT showed different self-regulation processes from learners supported by a human expert or a checklist, and the authors warned of metacognitive laziness, a dependence on AI that may hinder self-regulation. Darvishi et al. (2024) reported a randomised experiment on peer feedback in which students tended to rely on AI assistance and not learn from it. Self-regulation strategies filled part of the gap once the assistance was removed, but not as well as the assistance itself.

Self-regulation is not a fragile skill, and it can be taught. A meta-analysis of interventions in online and blended settings reported a moderate positive effect on achievement (ES = 0.69; Xu et al., 2023b). A scoping review of 163 studies found that most research did not address the preparatory and planning phases (Xu et al., 2023a). Syntheses of AI and self-regulation in higher education are accumulating (Banihashem et al., 2025; Lan & Zhou, 2025; Sardi et al., 2025). Design work has begun on adaptive scaffolds that turn real-time analytics into prompts (Li et al., 2025) and on generative AI feedback whose effect depends on how students perceive its source (Yilmaz et al., 2026). Work on conceptual frameworks that link AI capability, self-efficacy, and critical thinking awareness (Jia & Tu, 2024; Kong & Yang, 2024, the latter for K-12). Technology use also tracks strategy use more generally. In a study of 206 English-education students in Indonesia, more frequent use of digital tools correlated with metacognitive strategies ($r = .63$; Sakkir et al., 2025). That study concerns language learning and general digital tools, so it shows only that tool use and metacognitive strategies can go together, not that either causes learning. None of the works in this paragraph isolates university mathematics.

2.3. Fluency, offloading, and verification in mathematics

Two features of mathematics shape this question. First, chatbot output is fluent whether or not it is correct, so the student has to supply the judgement. Wang et al. (2025) call the risk epistemic offloading: the tool completes core mathematical work and the processes through which understanding forms are bypassed. They set this against uses that reduce anxiety and free attention for higher-level reasoning. Second, proof-based courses already carry a known risk of deference to authority. Yoon et al. (2024) noted that uncritical acceptance of AI arguments would extend a habit of relying on authority figures that earlier research had documented in proof learning. Verification therefore does double duty. It protects the student from errors, and it exercises the monitoring that self-regulation requires.

2.4. A coding lens for this review

To organise the charted evidence, the review uses three role labels for what the AI did in the learning process that a study describes. A co-thinker is an AI whose output the student treats as a claim to examine, test, and rewrite. An adaptive scaffold is support built or arranged for a specific phase of learning, such as prompts or a purpose-built chatbot. A cognitive crutch is an AI that performs the step the student was meant to perform, so that performance without the tool falls short of performance with it. These labels are interpretive devices introduced for this review. They are not validated constructs, and one study can show more than one role. The method section describes how they were applied.

3. Research Method And Materials

3.1. Design and reporting

The review followed the scoping-study framework of Arksey and O'Malley (2005) with the updated JBI guidance (Peters et al., 2020) and used the PRISMA-ScR checklist (Tricco et al., 2018) as a guide to reporting. A scoping design fits the aim, which is to map a small and heterogeneous body of evidence and to locate gaps, not to estimate an effect. The review

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

was not registered. It is not a systematic review of a bibliographic database, and its findings should be read accordingly.

3.2. Information sources and search

The search did not use Scopus or Web of Science. It was run in November 2026 through a general web search engine, followed by opening the publisher, repository, or preprint page of each candidate (arXiv, Springer, Research in Learning Technology, F1000Research, the Journal of Pedagogical Research, and institutional repositories). Roughly two dozen queries combined three groups of terms: the tool (ChatGPT, generative AI, chatbot, Copilot), the setting (mathematics, calculus, linear algebra, proof, abstract algebra, pre-service mathematics teachers), and the learning process (self-regulated learning, metacognition, verification, trust, reliance, offloading). Results from 2023 onward were examined. A commercial engine ranks the results and some publisher pages show only abstracts, so the search cannot be repeated exactly and does not claim to be exhaustive.

3.3. Eligibility criteria

A study was eligible if it (a) reported empirical data, or the design and validation of a learning tool with data; (b) took place in a university mathematics setting, including mathematics courses for pre-service teachers; (c) involved a generative AI chatbot used by students; (d) reported evidence on self-regulation, metacognition, verification, trust, reliance, or offloading, or on learning outcomes interpreted in relation to reliance; and (e) was published in English between 2023 and 2026. Studies were excluded if their samples came from schools, if they reported no findings, if the AI tool served instructors only, or if the setting was not specific to mathematics. When a study mixed school and university samples, only the university analysis was used.

3.4. Selection

One reviewer made all selection decisions. Table 1 reports the 21 records for which a decision and a reason were documented in the screening log (Supplementary File S1). Records dismissed at title level are not itemized, so Table 1 is a record of decisions and not a flow diagram with a reproducible total. Figure 1 borrows the layout of the PRISMA 2020 flow diagram (Page et al., 2021) but shows only the documented decisions and says so in its first box.

Table 1: Screening Decisions for the 21 Documented Records

Decision	Reason	Records
Included	Met all eligibility criteria (see Table 2)	12
Excluded	Sample drawn from schools, not universities	4
Excluded	No findings reported at the time of screening	1
Excluded	Sample not specific to mathematics	1
Excluded	University level or bibliographic details could not be confirmed	3
Total		**21**

Note: Records are those for which a decision and a reason are documented in Supplementary File S1. Records dismissed at title level are not itemized.

Twelve of the 21 records were included. Four were excluded because their samples came from schools, which is the most common reason in the log. Three were excluded because the university level or the bibliographic details could not be confirmed from the record, so the corpus may undercount eligible work.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

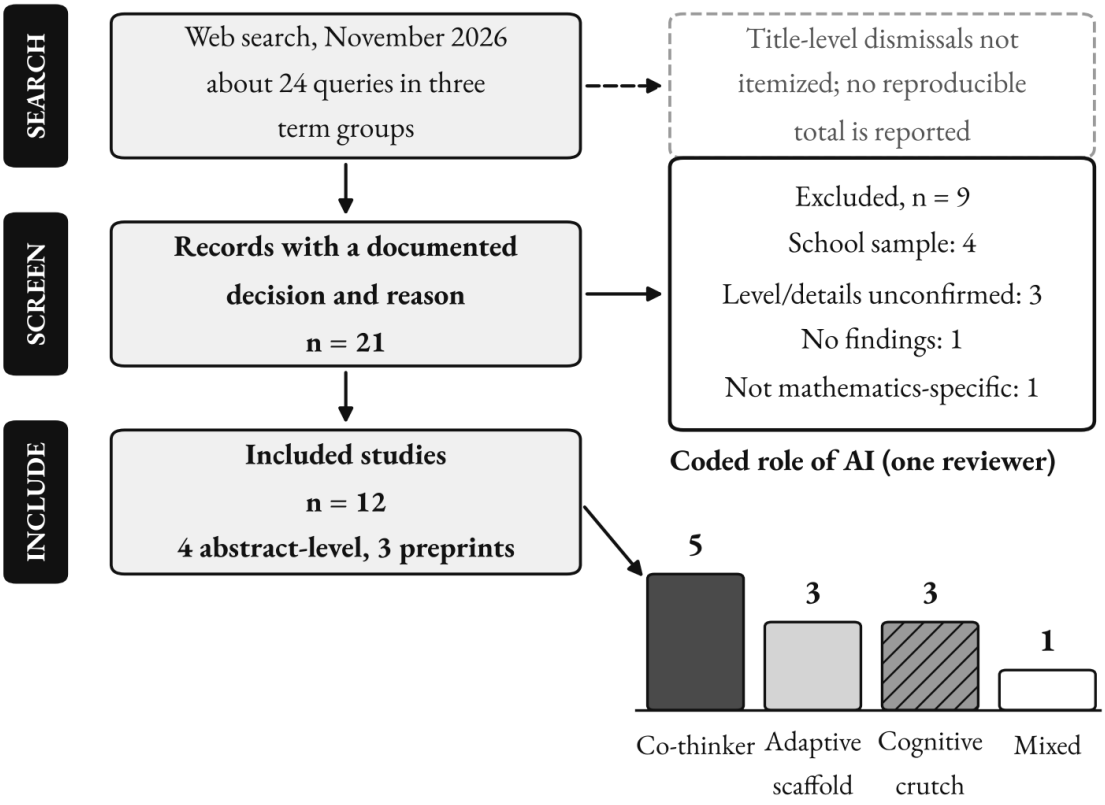


Figure 1: Study Selection Record and Coded Role of AI in the Twelve Included Studies

Figure 1 also previews the coding reported later. The largest group, five studies, is coded as co-thinker, and the three cognitive-crutch studies all rest on outcome or platform data and not on observed behaviour.

3.5. Charting and coding

For each included study the reviewer charted the country, course, design, sample, AI tool, learning-process focus, main findings, and the limitations stated or visible in the report. Four studies were charted from abstract-level records (Bakó & Szőke, 2026; Dello Iacono & Pepkolaj, 2026; Filiz & Gür, 2025; Rismanchian et al., 2026), which limits the detail available for them. The role label described in the literature review was assigned by the same reviewer after charting. No second coder was used, so no agreement statistic is reported. Critical appraisal is optional under PRISMA-ScR (Tricco et al., 2018). In its place, Table 3 lists the design features that limit each study's evidence.

4. Results And Discussion

4.1. Characteristics of the included studies

The twelve included studies appeared in 2024 (one study), 2025 (seven), and 2026 (four) and came from at least seven countries (Table 2). Six were small qualitative or mixed-methods studies with between three and 42 participants. One was a survey analysed with structural equation modelling. Three analysed performance or platform data without observing AI use directly. One was an experiment with 342 students, and one reported the design and validation of a chatbot. Courses ranged from calculus and linear algebra to abstract algebra, topology, and teacher education. Three studies were preprints in the versions read for this review (Klawa et al., 2025; Rismanchian et al., 2026; Wang et al., 2025).

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 2: Characteristics of the Included Studies

ID	Study	Setting and course	Design and sample	AI tool and focus
S1	Yoon et al. (2024)	South Korea; freshman mathematics education majors; number theory and advanced calculus statements	Exploratory task-based interviews; 3 students	ChatGPT 3.5; proving and evaluating output
S2	Agad et al. (2025)	Philippines (from author affiliations); senior BS Mathematics students; projective geometry proofs	Qualitative: proof task, ChatGPT feedback, revision, interview; 6 or 7 students (both reported)	ChatGPT, version not given; trust and revision
S3	Asare & Boateng (2025)	Ghana; pre-service mathematics teachers	Cross-sectional survey with structural equation modelling; planned n = 329; final N not reported	ChatGPT; usage, self-awareness, self-regulation, self-efficacy
S4	Bognár & Joós (2025)	Hungary; Mathematics 1, University of Dunaújváros	Observational comparison; 140 valid tests, 22 self-reported users; online 45-minute test	ChatGPT; scores and completion time
S5	Filiz & Gür (2025)	Prospective mathematics teachers; country not confirmed	Mixed methods; survey n = 42, interviews n = 4	ChatGPT; metacognitive awareness and problem solving
S6	Klawa et al. (2025)	United States; Modern Algebra I and II, Topology I	Survey (19 responses from 17 students, about 29% response) and 4 interviews	Microsoft Copilot; use and perceived accuracy
S7	Wang et al. (2025)	China; Wenzhou-Kean University; Business Calculus, Linear Algebra, Calculus III	Observational; unproctored homework versus proctored exams; sections of about 25 to 50 students	Use not tracked; epistemic offloading
S8	Zhuang (2025)	United States; Calculus I at a state university	Two-case qualitative study within a class of 12; surveys, six labs, chat histories	ChatGPT 3.5; perceptions and use
S9	Bakó & Szőke (2026)	Hungary; University of Debrecen; business mathematics	Experiment on differentiation tasks; 342 undergraduates; cluster analysis	ChatGPT-4; efficiency, motivation, critical checking
S10	Dello Iacono & Pepkolaj (2026)	Albania; first-year Calculus 2	In-class task, home comparison with ChatGPT, metacognitive questionnaire; 9 students	ChatGPT; metacognitive activity
S11	Rismanchian et al. (2026)	ALEKS platform data; college subsample used	Quasi-experimental panel of about 3.2 million interactions; proctored versus non-proctored	Before and after ChatGPT release; use not observed
S12	Rizki et al. (2026)	Indonesia; Universitas Pendidikan Indonesia; Differential Calculus	Design and validation; 3 experts and an 11-student pilot	Calcyon chatbot; self-regulation and computational thinking scaffolds

Note: Studies are ordered by year and author. Country is taken from the report unless stated.

The spread of designs means that no two studies measure the same thing. Participant counts range from 3 to 342 in studies with human participants, and one study works at the scale of millions of platform interactions (Rismanchian et al., 2026). Comparison across studies is therefore descriptive.

4.2. How students used AI output (RQ2)

Studies in which students examined AI output closely came mainly from proof-based and small-group settings. In the interview study by Yoon et al. (2024), three students doubted the correctness of ChatGPT's arguments, checked them, and rewrote the output in their own notation, and the authors argued that students should keep authority over their proofs. In three proof-based courses, Klawa et al. (2025) received 19 survey responses from 17 students. Eight of the 19 respondents did not use the tool although course policy allowed it, and users generally found it more useful for finding information than for completing proofs. In a Calculus 2 class of nine students, Dello Iacono and Pepkolaj (2026) reported

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

that comparing one's own solution with a ChatGPT solution raised monitoring, evaluation, and elaboration, while no student showed planning activity. Agad et al. (2025) followed senior mathematics students who revised geometric proofs after ChatGPT feedback. Most students trusted the feedback little, adopted suggestions selectively, and made revisions that were mostly communicative and left the core logic alone. The paper gives the number of participants as six in the abstract and seven in the methods.

Two cautions apply to this evidence. The first is that verification was often required. Klawa et al. (2025) note that the course asked students to justify the correctness of AI output, which may have raised the caution that students reported. The second is that required verification is not the same as accurate verification. In the calculus labs studied by Zhuang (2025), students had to judge each ChatGPT solution as correct, partly correct, or incorrect. The instructor observed that many of these judgements were wrong or missed errors, although the observation was not measured systematically. The two students followed in depth behaved differently. The student with a positive view of ChatGPT held 54 interactions and asked 26 follow-up questions about reasons. The skeptical student held 38 interactions and usually stopped once a correct solution appeared. The same assignments produced different behaviour, and the author attributes the difference to perception.

Evidence for offloading came from outcome data. Wang et al. (2025) compared unproctored homework with proctored exams in three courses across two semesters at one Chinese university. Where homework means were reported they stood between 96.6 and 99.1, while final-exam means ranged from 48.7 to 79.2 across sections, and homework-final correlations ran from −0.176 to 0.546. Rismanchian et al. (2026) used platform data and reported that time on AI-susceptible problems among college students fell by 2.8% per quarter after ChatGPT's release, about 26.9% over eleven quarters. The decline disappeared under proctoring, and the odds of a correct answer on proctored retention items fell by 25% cumulatively. Bognár and Joós (2025) reported that students who said they used ChatGPT in an online Mathematics 1 test scored a mean of 12.36 against 17.356 for non-users ($t = 4.04$, $p < .0001$). Completion time did not differ (23.36 versus 24.88 minutes, $p = .622$). In that study a Mood's median test gave $p = .064$, and use was self-reported, so the difference cannot be traced to ChatGPT alone.

Evidence for scaffolding was weaker still. In a survey of pre-service mathematics teachers in Ghana, Asare and Boateng (2025) reported that ChatGPT usage was associated with self-regulatory learning (standardized estimate .339), self-awareness (.229), and self-efficacy (.191). Self-regulatory learning and self-awareness partly mediated the link to self-efficacy (indirect estimates .110 and .082). The design is cross-sectional and self-reported, so the mediation describes a statistical model and not a demonstrated sequence. Filiz and Gür (2025) surveyed 42 prospective mathematics teachers and interviewed four, and concluded that ChatGPT can improve metacognitive awareness and problem-solving strategies through rapid feedback. The abstract-level record gives no numerical results. Rizki et al. (2026) built a calculus chatbot around four self-regulation components and validated it with three experts (mean Aiken's V of 0.92) and an 11-student pilot. They state that their results do not establish effectiveness. Bakó and Szőke (2026) found that ChatGPT-4 solved differentiation tasks with 94% accuracy and that 90% of 342 students performed well. Three learner clusters differed in engagement, and the abstract-level record does not say how performance was measured.

Figure 2 collects the numbers that carry the most weight in the preceding paragraphs. Panel A shows the gap between self-reported users and non-users in one mid-term test. Panel B plots the decline in time on task that follows from the reported rate of 2.8% per quarter, which the study itself reports as 26.9% over eleven quarters. Panel C shows the distance between unproctored homework means and proctored final-exam means, drawn as ranges because the study reports ranges across sections. Panel D shows the standardized estimates from the Ghanaian survey. The panels use different units and designs, so they should be read side by side and not compared on one scale.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

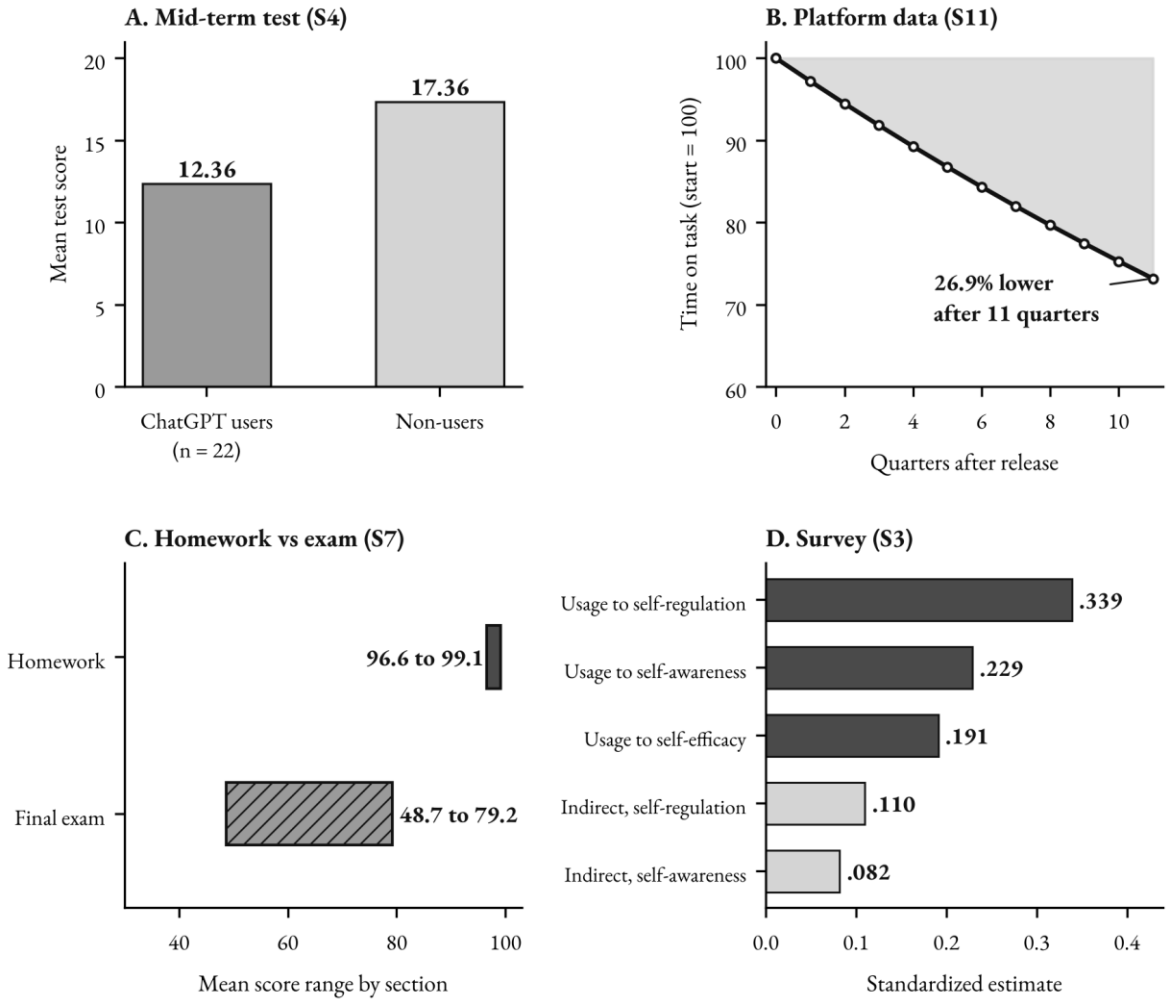


Figure 2: Quantitative Signals Reported by Four of the Included Studies

Table 3: Main Findings, Coded Role of AI, and Limits of the Evidence

ID	Main finding	Coded role	Limits of the evidence
S1	Students doubted correctness, checked arguments, and rewrote output; conceptions of proof, conceptions of AI, and ethics shaped use	Co-thinker	Three students, non-random sample; proof quality not evaluated
S2	Low trust in feedback; selective adoption; revisions mostly communicative, not changes to core logic	Co-thinker	Self-described behaviour only; participant count inconsistent; venue details incomplete
S3	Usage associated with self-regulation (.339), self-awareness (.229), self-efficacy (.191); indirect effects .110 and .082	Adaptive scaffold	Cross-sectional self-report; final N not reported; planned 329 versus 302 selected
S4	Users scored lower (12.36 vs 17.356, t = 4.04); completion time similar (23.36 vs 24.88 min, p = .622)	Cognitive crutch	Self-reported use, non-random groups; Mood's median test p = .064; short notice that ChatGPT was allowed; one test analysed
S5	Authors conclude ChatGPT can improve metacognitive awareness and problem-solving strategies through rapid feedback	Adaptive scaffold	Abstract-level record; no numerical results; DOI not listed
S6	8 of 19 respondents did not use AI; users found it better for information than for proofs; checking was a course requirement	Co-thinker	Small, non-anonymous, self-report; required checking may raise reported caution; one tool
S7	Homework means 96.6 to 99.1 against final means 48.7 to 79.2; homework-final r from −0.176 to 0.546	Cognitive crutch	No AI-use data; homework only a proxy; one institution; preprint

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

ID	Main finding	Coded role	Limits of the evidence
S8	Student with positive view: 54 interactions, 26 reasoning follow-ups; skeptical student: 38, usually stopped at a correct answer	Mixed	Two cases; labs graded on completeness; ChatGPT 3.5; one semester
S9	ChatGPT-4 solved differentiation tasks with 94% accuracy; 90% of students performed well; three learner clusters differed in engagement	Co-thinker	Abstract-level record; self-reported feedback; one institution
S10	Comparing own and ChatGPT solutions raised monitoring, evaluation, and elaboration; no planning activity	Co-thinker	Nine students; no numerical results; abstract-level record
S11	College time on AI-susceptible problems fell 2.8% per quarter (26.9% over 11 quarters), absent under proctoring; odds of a correct proctored answer fell 25%	Cognitive crutch	Observational platform data; use not observed; preprint; abstract-level record
S12	Expert validation grand mean 3.75 (mean Aiken's V 0.92); pilot grand mean 3.19 (79.67%)	Adaptive scaffold	Three experts; 11 pilot students, all female, one institution; no effectiveness data

Note: Roles were coded by one reviewer (see Charting and coding). Study IDs refer to Table 2.

Table 3 shows five co-thinker studies, three adaptive-scaffold studies, three cognitive-crutch studies, and one mixed case. These counts describe the twelve studies that were found and the reviewer's coding. They do not estimate how common each pattern is among students.

4.3. Strength of the evidence and gaps (RQ3)

A pattern runs through Table 3. Evidence of critical engagement comes from studies in which students were asked or required to evaluate AI output and in which behaviour was observed or self-reported. Evidence of offloading comes from studies that measured outcomes without observing use. None of the twelve studies links observed AI-use behaviour to unassisted performance in the same students. Until one does, students who check the output and students whose exam scores lag cannot be placed in the same cohort, and the two bodies of evidence cannot be reconciled.

The three phases of the cycle are unevenly covered. Performance-phase behaviours such as checking and follow-up questioning appear in several studies. Forethought appears in two: one study found no planning activity (Dello Iacono & Pepkolaj, 2026) and one designed planning support without testing it (Rizki et al., 2026). Only Asare and Boateng (2025) used a self-regulation scale, and three of its items were retained after factor analysis. The other studies infer self-regulation from behaviour or outcomes.

Reporting quality varies. Asare and Boateng (2025) give a planned sample of 329, a selection count of 302, and no final analysed sample. Agad et al. (2025) give two different participant counts. Wang et al. (2025) state that the gap exceeded 48 points in all three sections of one course, whereas the printed figures imply a smaller gap, 42.5, for one section. Three of the twelve studies are preprints. These points do not invalidate the studies, but they limit the weight that a single result can carry.

4.4. A testable model

The pattern of evidence suggests a model that is stated here as a set of hypotheses and not as a finding (Figure 3). Each phase of the cycle is paired with the evidence observed, the main gap, and a hypothesis that a later study could test. H1 states that a planning prompt before the first query raises later unassisted performance relative to unprompted use. H2 states that verification graded for accuracy narrows the gap between assisted and unassisted performance more than verification graded for completion, a distinction suggested by the Zhuang (2025) labs, which were graded on completeness. H3 states that asking students to predict their unassisted score before a proctored retrieval task narrows the gap between homework and exam performance reported by Wang et al. (2025) and Rismanchian et al. (2026). Testing these hypotheses requires usage logs linked to proctored retrieval for the same students, which none of the twelve studies had.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/mathscience5482

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

	Forethought *goal setting*	**Performance** *monitoring, checking*	**Self-reflection** *evaluation*
Evidence	No planning seen (S10); planning support designed but untested (S12).	Output checked when required (S1, S6, S8); time and exam gaps (S7, S11).	Own versus AI solution raised monitoring (S10); users scored lower (S4).
Main gap	Goal setting with AI is not measured.	Required checking was not always accurate (S8).	Predicted versus actual performance not measured.
Hypothesis (untested)	H1: A planning prompt before the first query raises unassisted performance.	H2: Checking graded for accuracy narrows the unassisted gap.	H3: Predicting one's score before proctored retrieval narrows the gap.

Figure 3: Evidence Map and Testable Hypotheses by Phase of the Self-Regulation Cycle

Figure 3 shows that the performance phase carries most of the evidence and the forethought phase carries the least. The hypotheses are ordered by how little is currently known, so H1 is the most speculative of the three.

4.5. Implications for teaching

Three implications follow, each tentative. First, if verification is assigned, its accuracy needs to be checked, because the instructor in Zhuang (2025) found many student judgements wrong. Second, homework scores alone cannot show whether learning occurred, given the large homework-exam gaps reported by Wang et al. (2025). Proctored retention items, as used by Rismanchian et al. (2026), give a second reading. Third, a short planning step before AI use costs little to add and addresses the least studied part of the cycle. A study of HyFlex mathematics learning made a related recommendation, that educators support self-regulation with guided learning plans and cognitive prompts (Parwati et al., 2025), and the same study reported a strong positive relationship between engagement and learning outcomes.

The design of the tool matters as much as the rule attached to it. In the high school experiment, a tutor version with safeguards largely removed the harm that unrestricted access caused (Bastani et al., 2025), which is the kind of contrast H2 proposes for verification. Feedback timing is another lever. Fakhri et al. (2024) studied the opposite condition, delayed feedback in programming, and found that behavioural and emotional engagement improved problem solving and helped students manage the delay. If so, instant AI feedback removes a delay that engagement used to absorb, although this inference needs testing in mathematics. Institutional conditions also count. A study of AI in university English speaking practice concluded that success depends on leadership, teacher training,

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

and policy, not on the technology alone (He et al., 2025). A study of online readiness found that infrastructure influenced student readiness more than teacher self-efficacy or attitude (Mappalotteng et al., 2025). Neither study concerns mathematics, so both are cited as context for adoption and not as evidence about learning.

The next studies also need a different way of observing students. Eye-tracking research on mathematical creativity often pairs gaze measures with stimulated recall interviews, which capture both visual behaviour and what participants say they were thinking (Farman et al., 2025). The same pairing of logged AI use with stimulated recall would let a study see what students do with a chatbot answer and how they explain it afterwards.

4.6. Limitations of the review

One reviewer selected, charted, and coded the studies, and no agreement statistic is available. The search used a web engine and publisher pages, not a bibliographic database, so relevant studies may be missing and the search cannot be repeated exactly. Four studies were charted from abstract-level records. The twelve studies differ in design, course, and outcome, which limits the review to a narrative synthesis. The role labels are interpretive, and a different coder could assign some studies differently. Three included studies are preprints whose later versions may change. Only English-language reports were considered, and three records could not be confirmed for eligibility and were left out.

5. Conclusion

This scoping review mapped twelve studies, published between 2024 and 2026, on generative AI and self-regulation in university mathematics. For RQ1, the studies are few, small or indirect, and spread across calculus, linear algebra, proofs, and teacher education. For RQ2, students examined AI output critically when the course or the researcher required it, and the clearest signs of offloading come from outcome data that show near-ceiling homework alongside weaker proctored performance. For RQ3, the evidence is thinnest where it matters most. No study connects observed use to unassisted performance in the same students, and forethought is almost unobserved.

The practical message is modest. Requiring verification, grading its accuracy, and checking learning with proctored retrieval items are changes that the evidence marks as worth testing. It does not show that they work. The contribution of the review is a map and three hypotheses, and its limits are those of a single-reviewer scoping search without a bibliographic database. A systematic review based on Scopus or Web of Science exports, two independent reviewers, and a registered protocol is the natural next step. Studies that log AI use, measure all three phases directly, and follow the same students to an unassisted test would answer the question that the present evidence leaves open.

References

Agad, A., Rocena, L. A., Cabalo, J. M. S., & Victorio, C. N. (2025). Revising with AI: Undergraduate students' perceptions and trust in AI-generated feedback on geometric proofs. *ATCM 2025 electronic proceedings*. https://atcm.mathandtech.org/EP2025/regular/22319.pdf

An, Q., Koh, J. H. L., & Liu, Q. (2026). Generative artificial intelligence in higher education: A systematic review of student use and learning outcomes. *Australasian Journal of Educational Technology, 42*(3), 62–82. https://doi.org/10.14742/ajet.10561

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Asare, B., & Boateng, F. O. (2025). Self-awareness and self-regulatory learning as mediators between ChatGPT usage and pre-service mathematics teacher's self-efficacy. *Journal of Pedagogical Research, 9*(2), 38–54. https://doi.org/10.33902/JPR.202530637

Bakó, M., & Szőke, S. (2026). Generative AI in higher education: Efficiency, motivation and challenges in teaching business mathematics. *Research in Learning Technology, 34*. https://doi.org/10.25304/rlt.v34.3673

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Banihashem, S. K., Bond, M., Bergdahl, N., Khosravi, H., & Noroozi, O. (2025). A systematic mapping review at the intersection of artificial intelligence and self-regulated learning. *International Journal of Educational Technology in Higher Education, 22*(1), Article 50. https://doi.org/10.1186/s41239-025-00548-8

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bognár, L., & Joós, A. (2025). The impact of allowing ChatGPT on student scores and test completion time in a mid-term math exam. *Dunakavics, 13*(2), 11–24. https://doi.org/10.63684/dk.2025.2.02

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00411-8

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

Dello Iacono, U., & Pepkolaj, L. (2026). ChatGPT as a student metacognitive reflection tool for recognizing mistakes in mathematics at university level. *Mathematics Teaching-Research Journal, 18*(3), 318–334.

Fakhri, M. M., Ahmar, A. S., Rosidah, R., Fadhilatunisa, D., & Tabash, M. (2024). Barriers to effective learning: Examining the influence of delayed feedback on student engagement and problem solving skills in ubiquitous learning programming. *Journal of Applied Science, Engineering, Technology, and Education, 6*(1), 69–79. https://doi.org/10.35877/454RI.asci2825

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

Farman, Siswono, T. Y. E., Lukito, A., Dewi, R. S., Ndayizeye, O., & Hali, F. (2025). Mathematical creativity: A systematic review of current research on eye-tracking technology. *Journal of Applied Science, Engineering, Technology, and Education, 7*(3), 451–464. https://doi.org/10.35877/454RI.asci4167

Filiz, A., & Gür, H. (2025). Students' perceptions and applications of metacognitive awareness levels in problem solving with ChatGPT. *Educational Process: International Journal, 14*(1), 1–29.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Green, K. H. (2024). Does ChatGPT know calculus? *Journal of Humanistic Mathematics, 14*(1), 248–257. https://doi.org/10.5642/jhummath.ILCN1025

He, P., Hanim A Hamid, A., & Mohd Nor, M. Y. (2025). Instructional leadership in the age of AI: Guiding effective use of AI tools in university English speaking practice. *Journal of Applied Science, Engineering, Technology, and Education, 7*(3), 564–580. https://doi.org/10.35877/454RI.asci4270

Jia, X.-H., & Tu, J.-C. (2024). Towards a new conceptual model of AI-enhanced learning for college students: The roles of artificial intelligence capabilities, general self-efficacy, learning motivation, and critical thinking awareness. *Systems, 12*(3), Article 74. https://doi.org/10.3390/systems12030074

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Klawa, H., Rajpal, S., & Thomas, C. (2025). *Gen AI in proof-based math courses: A pilot study* (arXiv:2509.13570v1). arXiv. https://arxiv.org/abs/2509.13570

Kong, S.-C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K–12 settings. *IEEE Transactions on Learning Technologies, 17*, 1562–1573.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Lan, M., & Zhou, X. (2025). A qualitative systematic review on AI empowered self-regulated learning in higher education. *npj Science of Learning, 10*(1), Article 21. https://doi.org/10.1038/s41539-025-00319-0

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1121). ACM. https://doi.org/10.1145/3706598.3713778

Li, T., Nath, D., Cheng, Y., Fan, Y., Li, X., Raković, M., Khosravi, H., Swiecki, Z., Tsai, Y.-S., & Gašević, D. (2025). Turning real-time analytics into adaptive scaffolds for self-regulated learning using generative artificial intelligence. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge*. ACM. https://doi.org/10.1145/3706468.3706559

Mappalotteng, A. M., Nurhikmah, H., Wijaya, H., Weismann, I. T. J., & Nur, I. A. F. M. (2025). Investigation readiness of teachers and students in online learning: A study quantitative method. *Journal of Applied Science, Engineering, Technology, and Education, 7*(3), 533–545. https://doi.org/10.35877/454RI.asci4212

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE, 19*(5), Article e0304013. https://doi.org/10.1371/journal.pone.0304013

Parwati, N. N., Suharta, I. G. P., Sudatha, I. G. W., Prameswara, P. B., & Halili, S. H. (2025). Optimization student learning outcomes through Hyflex learning based on student engagement. *Journal of Applied Science, Engineering, Technology, and Education, 7*(3), 427–440. https://doi.org/10.35877/454RI.asci4134

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis, 18*(10), 2119–2126. https://doi.org/10.11124/JBIES-20-00167

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E., & Kurd-Misto, E. (2026). *Faster completion, less learning: Generative AI reduced study time on math problems and the knowledge they build* (arXiv:2605.21629v3). arXiv. https://arxiv.org/abs/2605.21629

Rizki, L. M., Kusumah, Y. S., Jupri, A., Priatna, B. A., Aridanthy, V., Sowanto, & Nurlita, M. (2026). Calcyon: Design and validation of an AI-based chatbot framework integrating computational thinking and self-regulated learning in university calculus learning. *F1000Research, 15*, Article 1478. https://doi.org/10.12688/f1000research.180626.2

Sakkir, G., Atmowardoyo, H., & Dollah, S. (2025). The impact of digital learning technologies on students' English learning strategies in the new normal era. *Journal of Applied Science, Engineering, Technology, and Education, 7*(3), 465–476. https://doi.org/10.35877/454RI.asci4290

Sardi, J., Darmansyah, Candra, O., Yuliana, D. F., Habibullah, Yanto, D. T. P., & Eliza, F. (2025). How generative AI influences students' self-regulated learning and critical thinking skills? A systematic review. *International Journal of Engineering Pedagogy, 15*(1). https://doi.org/10.3991/ijep.v15i1.53379

Torres-Peña, R. C., Peña-González, D., Chacuto-López, E., Ariza, E. A., & Vergara, D. (2024). Updating calculus teaching with AI: A classroom experience. *Education Sciences, 14*(9), Article 1019. https://doi.org/10.3390/educsci14091019

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

Walkington, C. (2025). The implications of generative artificial intelligence for mathematics education. In *Proceedings of the forty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Kent State University. https://par.nsf.gov/biblio/10613898

Wang, S., Xia, Q., & Ye, Q. (2025). *Synthetic fluency and epistemic offloading in undergraduate mathematics in the age of AI* (arXiv:2512.21045v1). arXiv. https://arxiv.org/abs/2512.21045

Xu, Z., Zhao, Y., Liew, J., Zhou, X., & Kogut, A. (2023a). Synthesizing research evidence on self-regulated learning and academic achievement in online and blended learning environments: A scoping review. *Educational Research Review, 39*, Article 100510. https://doi.org/10.1016/j.edurev.2023.100510

Xu, Z., Zhao, Y., Zhang, B., Liew, J., & Kogut, A. (2023b). A meta-analysis of the efficacy of self-regulated learning interventions on academic achievement in online and blended environments in K-12 and higher education. *Behaviour & Information Technology, 42*(16), 2911–2931. https://doi.org/10.1080/0144929X.2022.2151935

Yilmaz, M., Temur, H. B., Emmungil, L., Çelik, E., Gauthier, A., & Cukurova, M. (2026). Supporting self-regulated learning through generative AI feedback in online higher education: The importance of student perceptions of the source of feedback. *International Journal of Educational Technology in Higher Education, 23*(1). https://doi.org/10.1186/s41239-026-00592-y

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education, 56*(7), 1531–1551. https://doi.org/10.1007/s11858-024-01629-0

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments, 11*, Article 28. https://doi.org/10.1186/s40561-024-00316-7

Zhuang, Y. (2025). Lessons from using ChatGPT in calculus: Insights from two contrasting cases. *Journal of Formative Design in Learning, 9*, 25–35. https://doi.org/10.1007/s41686-025-00098-2

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.