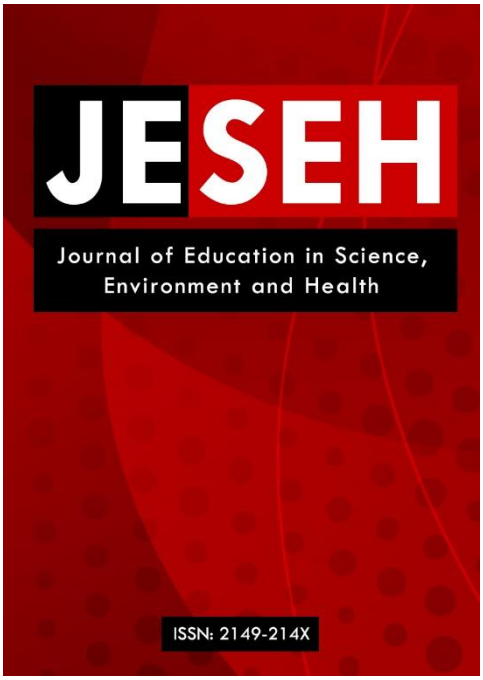

ISSN: 2149-214X

Journal of Education in Science, Environment and Health

www.jeseh.net

Conditional Effects of AI Homework Tools on Students' Academic Performance: A Systematic Synthesis of Empirical Evidence

Seyma Irmak[1], Kaan Bati[2]
[1]Amasya University
[2]Hacettepe University

To cite this article:

Irmak, S. & Bati, K. (2026). Conditional effects of AI homework tools on students' academic performance: A systematic synthesis of empirical evidence. *Journal of Education in Science, Environment and Health (JESEH), 12*(2), 160-173. https://doi.org/ 10.55549/jeseh.896

JESEH
Journal of Education in Science, Environment and Health
Volume 12, Issue 2, 2026
https://doi.org/10.55549/jeseh.896

Conditional Effects of AI Homework Tools on Students' Academic Performance: A Systematic Synthesis of Empirical Evidence

Seyma Irmak, Kaan Bati

Article Info

Article History

Published:
01 April 2026

Received:
27 January 2026

Accepted:
05 March 2026

Keywords
Artificial intelligence in education,
Generative AI,
Systematic narrative synthesis design

Abstract

The rapid diffusion of generative artificial intelligence (AI) tools into educational contexts has fundamentally transformed how students approach homework, academic writing, and independent learning tasks. Whilst AI-assisted homework tools promise efficiency, personalization, and immediate feedback, there remains some debate over their implications for academic performance and learning quality. The present study proffers a thorough synthesis of empirical evidence, examining how students' academic performance differs when using AI homework tools compared to traditional homework methods. The review draws on experimental, quasi-experimental, and observational research conducted across secondary and higher education contexts. The findings of the study indicate that AI homework tools are associated with significantly higher grades and writing scores in most controlled comparisons, particularly in language learning contexts, with effect sizes ranging from medium to large. However, the evidence also reveals important trade-offs, including reduced knowledge retention, lower originality, and diminished critical thinking in some settings. The synthesis demonstrates that AI tools primarily optimize output quality rather than learning processes, and that their effectiveness is highly conditional on task characteristics, assessment timing, implementation fidelity, and learner characteristics.

Introduction

Homework has long been considered a core component of formal education, functioning as a mechanism through which students rehearse skills, consolidate conceptual understanding, and develop independent learning habits (Cooper et al., 2006; Dettmers et al., 2009). Classical and contemporary research on homework effectiveness has consistently emphasized that its educational value depends not merely on the quantity of tasks assigned but, on their quality, alignment with instructional goals, and the nature of feedback provided (Hattie, 2009; Trautwein & Köller, 2003). As educational technologies have evolved, homework practices have been repeatedly reshaped, from paper-based assignments to online learning platforms and adaptive systems that offer automated feedback and progress monitoring (Dede, 2014).

The emergence of generative artificial intelligence (GenAI) represents a qualitative shift in this trajectory. Unlike earlier educational technologies that delivered pre-scripted content or rule-based feedback, contemporary AI systems—such as large language models (LLMs), automated writing evaluation tools, and intelligent tutoring systems—can generate explanations, examples, and complete textual outputs in response to students' prompts (Zhou et al., 2024). These systems increasingly accompany students during homework completion, offering real-time assistance that resembles aspects of human tutoring and raising fundamental questions about authorship, cognition, and learning responsibility (Smerdon, 2024).

The swift integration of AI-driven homework assistants into educational settings has ignited a rigorous scholarly discourse. Proponents highlight the potential of these tools to democratize high-quality feedback and provide vital support for learners facing language barriers or gaps in prior knowledge, primarily through personalized and immediate interventions (Song & Song, 2023; Tamimi et al., 2024). Within this framework, AI functions as a scalable mechanism for scaffolding (Wood et al., 1976), extending instructional guidance beyond the physical classroom and operationalizing Vygotsky's (1978) socio-constructivist principles in the digital age (Luckin et al., 2016). Critics, however, caution that reliance on AI may encourage surface-level engagement, undermine critical thinking, and weaken knowledge retention by offloading essential cognitive processes to automated systems (Yavich, 2025; Yang, 2025). Concerns regarding academic integrity, authorship, and educational equity further complicate the discourse (Smerdon, 2024).

Despite the prominence of these debates, empirical evidence remains fragmented. Individual studies report divergent findings, with some documenting substantial improvements in assignment quality and grades (Chen & Gong, 2025; Song & Song, 2023) and others finding negligible or even negative effects on learning-related outcomes such as retention and independent reasoning (Yang, 2025; Yavich, 2025). Moreover, academic performance is operationalized inconsistently across studies, ranging from immediate assignment scores to delayed retention measures and qualitative indicators of engagement (Kwak, 2025). Consequently, educators and policymakers face challenges in interpreting whether observed performance gains reflect genuine learning or merely improved outputs.

To resolve this critical ambiguity and distinguish cognitive growth from superficial task completion, the present study addresses this gap by synthesizing empirical studies that directly compare AI-assisted homework with traditional homework methods. Rather than treating AI effectiveness as a binary outcome, the analysis adopts a conditional perspective, examining how task complexity, assessment timing, implementation fidelity, and learner characteristics shape observed effects (AlShibli et al., 2025; Ward et al., 2025). The research question is: How do students' academic performance outcomes differ when using AI homework tools compared to traditional homework methods?

Literature Review

AI Homework Tools and Theoretical Perspectives

AI homework tools encompass a diverse spectrum of technologies designed to facilitate out-of-class learning, ranging from long-standing intelligent tutoring systems (ITS) and automated writing evaluation (AWE) tools to contemporary generative AI (GenAI) models (Kelly et al., 2013; Zhou et al., 2024). While traditional AI tools focused on adaptive feedback within structured environments, the latest generation is distinguished by its generative capacity, enabling the dynamic creation of explanations, summaries, and complex academic drafts that go beyond static content delivery.

From a theoretical standpoint, AI homework tools intersect with multiple learning theories. Behaviorist perspectives emphasize the role of immediate feedback and reinforcement, which AI systems can deliver consistently at scale (Skinner, 1961; Kulik & Fletcher, 2016). By providing instantaneous corrections and rewards, these tools mirror the programmed instruction model, ensuring that learners consolidate correct responses before advancing to more complex tasks. Cognitive load theory suggests that AI-generated explanations and worked examples may reduce extraneous cognitive load, allowing learners to allocate more cognitive resources to germane processing (Sweller et al., 2011, pp. 15-16, 102). In contrast, constructivist and socio-cognitive frameworks highlight the importance of active knowledge construction, metacognitive monitoring, and productive struggle (Kapur, 2008; Zimmerman, 2002). Within these frameworks, AI tools may function either as scaffolds that support learning or as substitutes that bypass essential cognitive engagement, depending on how they are integrated into instructional design (Zhou et al., 2024).

Homework, Performance, and Learning Depth

Academic performance is frequently assessed using proximal indicators such as assignment grades, rubric-based scores, or standardized test results (Hattie, 2009). While such measures provide tangible evidence of achievement, they do not necessarily capture deeper learning outcomes, including conceptual understanding, transfer, and long-term retention (Bjork & Bjork, 2011; Soderstrom & Bjork, 2015). Research on surface versus deep learning demonstrates that learners can achieve high immediate performance while developing fragile knowledge structures that do not support future application (Marton & Säljö, 1976).

In AI-assisted homework contexts, this distinction becomes particularly salient. Generative AI systems are highly effective at improving surface features of academic work, such as grammatical accuracy, coherence, and organization (Chen & Gong, 2025; Song & Song, 2023). However, their influence on higher-order outcomes—critical thinking, originality, and epistemic judgment—remains uncertain and empirically contested (Yavich, 2025; Smerdon, 2024). Evaluating AI homework tools, therefore, requires attention not only to performance gains but also to the nature and durability of learning they promote, distinguishing between cognitive offloading and genuine skill acquisition.

Empirical Evidence on AI, Engagement, and Cognition

Empirical evidence regarding AI's role in education underscores a complex relationship between engagement, performance, and cognitive processing. Recent studies suggest that AI tools can bolster student motivation by offering personalized, interactive learning experiences that cater to individual needs (Bognár & Khine, 2025; Tamimi et al., 2024). In terms of measurable outcomes, performance gains of approximately 15% to 35% have been reported, particularly when adaptive learning platforms and intelligent tutoring systems provide targeted scaffolding (Kwak, 2025; Ward et al., 2025). However, these gains are not universally distributed; evidence indicates that the impact on cognitive depth varies significantly based on demographic factors, prior achievement, and distinct patterns of use—ranging from active scaffolding to passive reliance (AlShibli et al., 2025).

Cognitive impacts of AI use are similarly mixed. Some studies report positive associations between AI-supported self-regulation and problem-solving (Zhou et al., 2024), while others caution that excessive reliance on AI may reduce students' willingness to engage in independent analysis and epistemic monitoring (Yavich, 2025). Engagement effects may also change over time, with initial enthusiasm diminishing as novelty effects fade (Bognár & Khine, 2025). Collectively, these findings underscore the importance of examining AI homework tools within broader theoretical frameworks of self-regulated learning and cognitive engagement (Zimmerman, 2002).

Methodology

Research Design

This study employed a systematic narrative synthesis design (Popay et al., 2006) to integrate empirical evidence comparing AI-assisted homework tools with traditional homework methods. A narrative approach was deemed most appropriate due to the extensive methodological and conceptual heterogeneity across primary studies, including variations in educational levels, disciplinary domains, and AI tool functionalities. Unlike meta-analytic techniques that necessitate strictly commensurable effect size reporting, narrative synthesis facilitates a theory-informed integration of diverse findings while preserving the contextual nuances essential for interpreting educational interventions (Popay et al., 2006). The review process followed the PRISMA 2020 guidelines to ensure transparency and replicability (Page et al., 2021), while the overall systematic framework was informed by principles of evidence-based education research (Gough et al., 2017).

Search Strategy

A systematic search was conducted using the Elicit research engine (https://elicit.com), selected for its advanced natural language processing (NLP) capabilities, which facilitate semantic searching. Unlike traditional keyword-matching databases, this approach identifies relevant literature based on conceptual meaning, ensuring comprehensive coverage of the rapidly evolving AI nomenclature (e.g., distinguishing between 'Generative AI', 'Large Language Models', and 'Intelligent Tutoring Systems'). The search leveraged the Semantic Scholar and OpenAlex databases, which collectively index over 200 million scholarly publications. Combinations of the following keywords were used: *artificial intelligence, generative AI, homework, academic performance, student achievement, writing, problem solving,* and *learning outcomes*. Filters were applied to peer-reviewed journal articles and refereed conference proceedings published between January 2013 and January 2025.

Study Selection Process

The study selection followed a rigorous two-stage screening process (See Figure 1 for the PRISMA flow diagram). In Stage 1 (Preliminary Screening), titles and abstracts were independently screened against the research objective. Studies were excluded if they were purely theoretical, focused exclusively on teacher-facing AI, or did not involve formal educational settings. In Stage 2 (Full-Text Eligibility Assessment), the remaining publications underwent a comprehensive full-text review. The primary reason for the subsequent exclusion of 40 studies was the absence of a rigorous comparative condition. This systematic refinement resulted in a final corpus of 10 empirical studies that met all criteria. To enhance reliability, the selection was finalized through an iterative review process to ensure perfect alignment with the research question.

Inclusion and Exclusion Criteria

Studies were included if they satisfied the following conditions: (a) AI-powered tools were used directly by students during homework completion; (b) participants were enrolled in formal secondary or tertiary education; (c) academic performance outcomes were reported using quantitative or mixed-methods measures; (d) the study employed an empirical research design (experimental, quasi-experimental, or observational); and (e) a clearly defined comparison condition using traditional, non-AI homework methods was present. Studies were excluded if they were limited to conceptual discussions, opinion pieces, or examined AI use solely during supervised in-class activities.

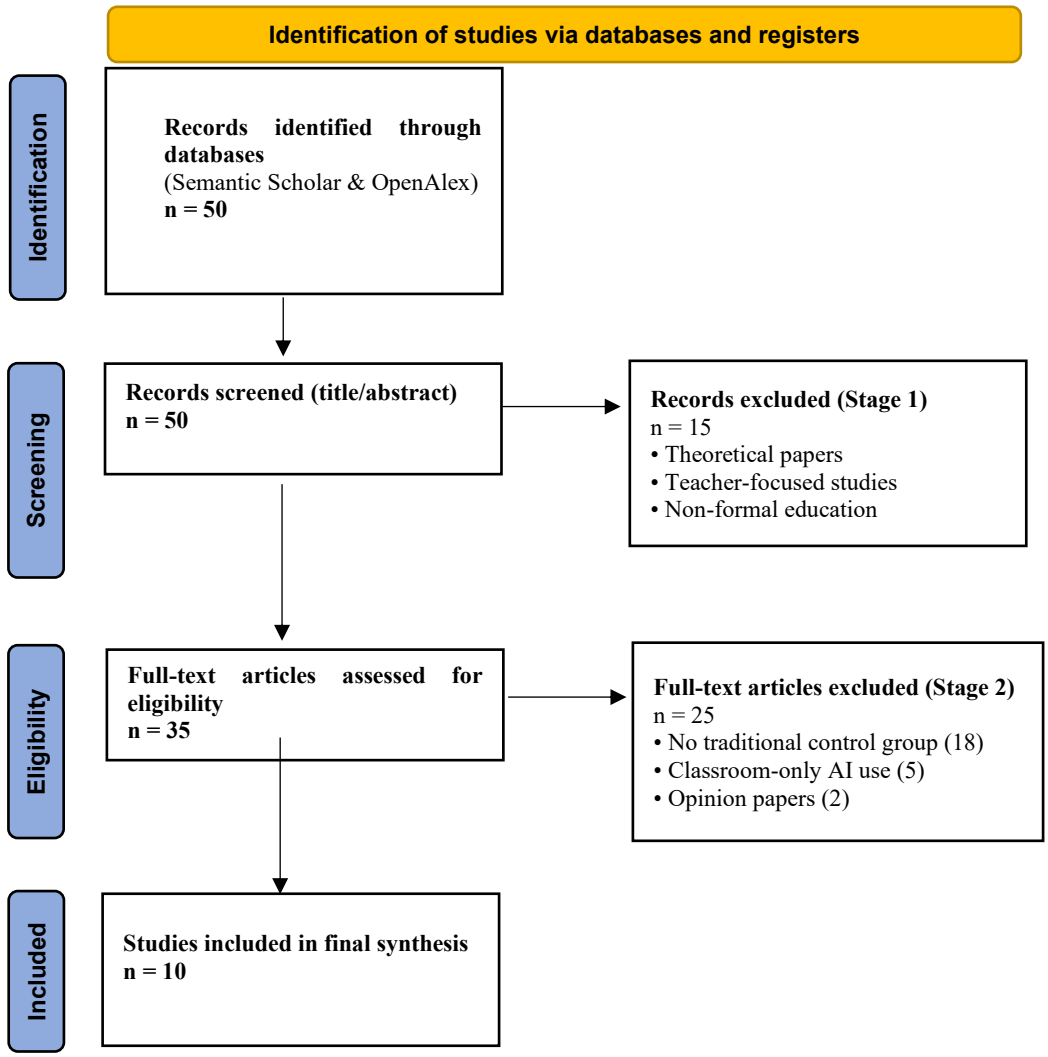

Figure 1. Centre the caption below the figure

Data Extraction and Quality Appraisal

A structured data extraction protocol was applied to record educational contexts, sample sizes, AI tool types, and performance metrics. Where effect sizes (Cohen's d) were not explicitly provided, the direction and magnitude of effects were inferred from reported statistical results. To ensure methodological rigor, a narrative critical appraisal framework was utilized (Petticrew & Roberts, 2006). Studies were assessed for internal validity, transparency of outcome measures, and the adequacy of their comparison conditions. Studies with robust experimental controls were given greater interpretive weight during the synthesis.

Data Synthesis

Findings were integrated using a thematic narrative synthesis approach (Popay et al., 2006). Results were organized according to: (1) disciplinary domain, (2) assessment timing (immediate performance vs. delayed retention), and (3) learner characteristics. This analytical strategy enabled the identification of conditional effects—circumstances under which AI-assisted homework proves beneficial, neutral, or detrimental—thereby avoiding overgeneralized conclusions regarding AI effectiveness.

Results

Overview of the Evidence Base

The synthesis drew on 10 empirical studies that explicitly compared AI-assisted homework practices with traditional approaches across secondary and higher education settings (See Table 1). Collectively, these studies provide a heterogeneous but informative evidence base regarding how AI tools influence academic performance under varying instructional conditions. However, a notable disciplinary divide was observed: while AI's impact is extensively documented in language-related tasks (e.g., EFL writing, translation), empirical evidence in STEM and pure science domains remains emerging and shows more nuanced results.

Table 1. Characteristics of included studies

Study	Study Design	Sample Size (AI / Traditional)	Duration	Educational Level	Subject Area	AI Tools Used	Traditional Method
Wale, 2024	Quasi-experimental (pretest–posttest)	92 total (groups not specified)	8 weeks	Undergraduate (third year)	EFL academic writing	Writerly, Google Docs	Paper–pencil feedback
Khaustov et al., 2024	Quasi-experimental	25/25	Not specified	Undergraduate (second year)	Translation studies	ChatGPT-4.0, Criterion, YandexGPT	Traditional instruction
Kelly et al., 2013	Quasi-experimental	63 total (8 classes /9 classes)	60-minute session	K–12 (grades 7–8)	Mathematics	ASSISTments (web-based)	Practice without feedback
Chen & Gong, 2025	Mixed methods, quasi-experimental	25/25	16 weeks	Undergraduate (third year)	Chinese as a second language	ChatGPT	Traditional teacher instruction
Smerdon, 2024	Observational	46/23 (survey respondents)	One semester	Undergraduate	Research proposal	Not specified	Not specified
Song & Song, 2023	Mixed methods, RCT	25/25	12 weeks	Undergraduate	EFL writing	ChatGPT	Traditional teacher instruction
Faridoon et al., 2025	Observational	Not specified	One semester	Higher education	Not specified	AI tutoring systems	Not specified
AlShibli et al., 2025	Quasi-experimental	32/32	4 weeks	Undergraduate (first year)	Computer science	Not specified	Textbooks, notes
Tamimi et al., 2024	Observational, mixed methods	115 total	One semester	High school	General homework	ChatGPT, StudyPool, TutorEva, OddityAI	Not specified
Boumediene & Bouakkaz, 2024	Observational survey	~1,200 students	Fall 2023 – Spring 2024	High school (grades 9–12)	Multiple subjects	ChatGPT	Not specified

Seven studies employed experimental or quasi-experimental designs with clearly defined control groups using traditional homework methods, while three relied on observational or correlational designs. Sample sizes ranged from small classroom-based interventions ($n \approx 50–80$) to large-scale institutional datasets exceeding 1,000 participants (Boumediene & Bouakkaz, 2024). The diversity of contexts allowed for cross-study comparisons of task types, disciplinary domains, and assessment practices.

Taxonomy of AI Tools Across Reviewed Studies

A critical prerequisite for interpreting the findings across the reviewed studies is recognizing that "AI homework tools" is not a monolithic category. The tools identified in the synthesis span two functionally distinct paradigms, each with fundamentally different mechanisms of action and pedagogical implications. The first category comprises *Intelligent Tutoring Systems (ITS)*, such as ASSISTments (Kelly et al., 2013), which operate within structured, domain-specific frameworks. These systems are designed primarily to *provide formative feedback*: they evaluate student responses against predefined correct answers, diagnose misconceptions, and adaptively guide learners through problem-solving sequences. Their primary function is to scaffold existing knowledge, not to generate new content on the student's behalf. The second category encompasses *Generative AI (GenAI) tools*, including large language models such as ChatGPT (Chen & Gong, 2025; Khaustov et al., 2024; Song & Song, 2023; Tamimi et al., 2024; Boumediene & Bouakkaz, 2024) and YandexGPT (Khaustov et al., 2024), as well as automated writing evaluation (AWE) platforms such as Writerly and Criterion (Wale, 2024; Khaustov et al., 2024).

It should be noted that Wale (2024) also references Google Docs as the writing environment in which the AWE tool was deployed; Google Docs itself is not an AI tool and is therefore excluded from this taxonomy. Unlike ITS, GenAI and AWE tools are designed for *direct content generation*: they can produce complete textual outputs, draft academic essays, translate texts, and provide holistic stylistic revisions in response to open-ended student prompts. This generative capacity fundamentally distinguishes them from feedback-oriented ITS. A third, less clearly defined category in the evidence base consists of *AI tutoring and homework assistance platforms* (e.g., TutorEva, OddityAI, StudyPool, and the unspecified AI tutoring systems referenced in Faridoon et al., 2025 and AlShibli et al., 2025), which blend elements of both paradigms by combining answer support with explanatory scaffolding.

Table 2. Quantitative performance results from studies with controlled comparisons

Study	Outcome Measure	AI Group Mean (*SD*)	Traditional Group Mean (*SD*)	Statistical Significance	Effect Size	Direction
Wale, 2024	IELTS writing post-test	54.68	45.85	$p < .05$	$d = 0.924$	AI better
Khaustov et al., 2024	Business letter writing	Not reported	Not reported	$p = .001$	Not reported	AI better
Khaustov et al., 2024	Contrastive/comparative essay	Not reported	Not reported	$p = .002$	Not reported	AI better
Khaustov et al., 2024	Argumentative essay	Not reported	Not reported	$p = .001$	Not reported	AI better
Khaustov et al., 2024	Film/book review	Not reported	Not reported	$p = .001$	Not reported	AI better
Kelly et al., 2013	Learning gains (pre–post)	Not reported	Not reported	$p = .1065$	$d = 0.56$ (WBH generally)	No significant difference
Chen & Gong, 2025	Academic writing post-test	89.74 (9.08)	82.15 (10.23)	$p < .05$	Not reported	AI better
Smerdon, 2024	Task performance	Not reported	Not reported	Not significant	< 1 mark / 100	No difference
Song & Song, 2023	Overall writing proficiency	59.12 (14.23)	45.18 (15.62)	$p < .001$	$d = 0.76$	AI better
Song & Song, 2023	Writing content	15.96 (3.71)	13.71 (3.12)	$p = .003$	$d = 0.65$	AI better
Song & Song, 2023	Writing organization	16.56 (3.54)	13.63 (3.63)	$p < .001$	$d = 0.84$	AI better
Song & Song, 2023	Language use	19.89 (4.82)	15.89 (4.12)	$p < .001$	$d = 0.88$	AI better
AlShibli et al., 2025	Overall average score	85.3%	79%	Not reported	Not reported	AI slightly better
AlShibli et al., 2025	Knowledge retention	20% demonstrated retention	Not reported	Not reported	Not reported	AI worse

This functional distinction is not merely taxonomic; it carries direct interpretive consequences for the reported outcomes. Studies employing ITS in mathematics contexts (Kelly et al., 2013) produced modest or non-significant effects on learning gains (d = 0.56, p = .1065), consistent with the ITS design goal of deepening procedural knowledge through targeted feedback. In contrast, studies utilizing GenAI tools in writing and language tasks reported substantially larger effects on output quality (e.g., d = 0.924 in Wale, 2024; d = 0.76–0.88 in Song & Song, 2023), which reflects the capacity of these tools to directly enhance the surface features of student-produced text. However, this performance advantage comes with a critical caveat: because GenAI tools can generate entire drafts, the observed gains may reflect the tool's output quality rather than the student's cognitive engagement. By contrast, ITS-driven gains, while more modest, are more directly attributable to student learning activity. Consequently, observed performance gains and their relationship to genuine cognitive engagement must be interpreted in light of the tool's fundamental purpose—whether it is designed to scaffold student thinking or to substitute for it.

Domain-Specific Outcomes: Language Proficiency vs. STEM Reasoning

A consistent pattern of high efficacy emerged in language-related domains (EFL, writing, translation). Studies reported statistically significant advantages for AI-assisted groups (Chen & Gong, 2025; Song & Song, 2023). As shown in Table 2, Wale (2024) reported a post-test IELTS writing score of 54.68 for the AI group compared to 45.85 for the traditional group (d = 0.924). Similarly, Song & Song (2023) documented large effect sizes across multiple dimensions, including writing organization (d = 0.84) and language use (d = 0.88). In contrast, findings from STEM and science-related domains were notably more mixed. Kwak (2025) and Kelly et al. (2013) found modest or non-significant differences in mathematics and problem-solving (p = .1065 in Kelly et al.). A striking disciplinary contrast is evident here: while AI excels in improving the "surface features" and organization of language tasks, its impact on the multi-step conceptual reasoning required in science and mathematics appears limited when the tool provides only answer verification rather than conceptual scaffolding.

The Performance-Learning Paradox: Immediate Gains vs. Retention

The temporal aspect of assessment was identified as a pivotal moderator of performance outcomes. Research focusing on immediate post-task results has consistently demonstrated a preference for AI-assisted homework (AlShibli et al., 2025; Song & Song, 2023). However, a "Performance-Learning Paradox" was identified: Proximal Success: AI users frequently attain higher grades in assignments (e.g., 85.3% vs. 79% in AlShibli et al., 2025).

Distal Failure: However, when delayed retention measures were employed, these advantages diminished. In the study conducted by Yang (2025), it was observed that students who placed significant reliance on artificial intelligence exhibited a substantially diminished capacity to retain conceptual knowledge. A seminal study by AlShibli et al. (2025) revealed a striking finding: while AI users exhibited superior report quality, only 20% demonstrated evidence of knowledge retention during independent assessments. This finding suggests that AI may encourage the development of "fragile knowledge" (Marton & Säljö, 1976), which is not conducive to long-term mastery.

Secondary Outcomes: The Trade-off Between Efficiency and Engagement

Beyond grades, AI-assisted students demonstrated higher time efficiency and motivation (Bognár & Khine, 2025; Ward et al., 2025). As detailed in Table 3, motivation gains were especially pronounced among students with lower prior achievement, who described AI tools as reducing frustration and cognitive overload (Tamimi et al., 2024).

However, the qualitative data revealed critical trade-offs. Smerdon (2024) and Yavich (2025) reported a decline in originality and independent reasoning. Instructors noted an increase in the uniformity of student work, with Boumediene & Bouakkaz (2024) observing that enhanced grammar and coherence were frequently accompanied by diminished critical thinking scores. A salient finding from Yavich's (2025) study was that less than 40% of students exhibited mastery of the content without AI assistance, thereby indicating a transition from cognitive engagement to "cognitive offloading."

Table 3. Secondary outcomes and quality indicators

Study	Completion Rates	Engagement / Motivation	Time Efficiency	Work Quality Dimensions
Wale, 2024	Not reported	Positive perceptions toward AI tools; increased motivation	Not reported	Improved task achievement, coherence, lexical resource, and grammar
Chen & Gong, 2025	Not reported	Writing motivation: AI M = 20.06 (SD = 3.33) vs. Traditional M = 18.21 (SD = 3.58), p = .001, d = 0.52	AI group completed reports faster	Enhanced ideas, coherence, lexicon, and grammatical range
Song & Song, 2023	Not reported	Higher participation, greater assignment consistency, stronger motivation	Not reported	Improved across all writing dimensions
Faridoon et al., 2025	Not reported	40% of teachers observed reduced effort	Not reported	Improved subject comprehension and problem-solving skills
AlShibli et al., 2025	Timely submissions increased from 75% to 85%	Not reported	Not reported	Better report quality but worse quiz performance
Boumediene & Bouakkaz, 2024	Not reported	Not reported	Not reported	Better grammar and flow, but lower originality and critical thinking

Note. Outcomes are reported as described in the original studies. Blank cells indicate that the indicator was not assessed or not reported.

Differential Effects by Learner Characteristics

The evidence presented indicates a correlation between the impact of AI and prior student achievement.

Strategic Use: High-achieving students employed AI tools to refine their ideas and check their understanding, achieving moderate gains without any loss of comprehension (Chen & Gong, 2025).

Dependency risk: Conversely, students with lower academic achievements exhibited a propensity to depend on direct AI-generated outputs, thereby augmenting their short-term scores while concomitantly escalating the likelihood of long-term cognitive dependency (AlShibli et al., 2025; Ward et al., 2025).

The findings, when considered collectively, reveal that AI-based homework tools do not consistently exert uniform effects on academic performance. Instead, the impact of AI in education is contingent on disciplinary context, assessment design, learner characteristics, and the degree to which AI use is pedagogically structured. These conditional patterns provide a necessary foundation for interpreting performance gains and inform the discussion of instructional implications.

Discussion

Reinterpreting Academic Performance: Productivity vs. Cognitive Change

The findings of this synthesis indicate that gains in academic performance associated with AI-assisted homework are fundamentally dependent on how performance is operationalized. Our results align with Gowtham et al. (2026) and Kwak (2025), suggesting that AI tools tend to optimize measurable outputs—such as assignment grades and completion rates—rather than the underlying cognitive processes. From a Behaviorist perspective, this optimization reflects Skinner's (1961) vision of immediate reinforcement, where students are rewarded for correct outputs. Quantitative indicators such as GPA improvements, increased completion rates, and higher standardized test scores have frequently been reported following the integration of AI tools, with gains ranging from approximately 15% to 35% in recent studies (Gowtham et al., 2026; Kwak, 2025). These results align with the present synthesis, which found significant performance advantages in language-focused assignments. However, as fundamentally argued by Soderstrom and Bjork (2015), such proximal indicators capture only a narrow slice of educational effectiveness. High performance on immediate tasks can often be a misleading proxy for actual learning, potentially obscuring significant declines in higher-order cognitive engagement and long-term retention. In the context of AI-assisted homework, this suggests that the 'polished output' may serve as a facade for what Lodge et al. (2023) describe as the erosion of assessment validity in the age of AI.

Surface Performance Gains Versus Deep Learning Outcomes

A central contribution of this study lies in clarifying the distinction between surface-level performance and durable learning outcomes. While AI tools demonstrably enhance linguistic accuracy and task efficiency, their impact on science process skills remains mixed. Our findings resonate with Zhai (2023), who argued that while generative AI can perform complex scientific tasks, it risks undermining the inquiry process if students use it as a surrogate for conceptual thinking rather than a supportive tool. This pattern is further supported by Kasneci et al. (2023), who emphasized in their comprehensive review that while large language models can enhance productivity, they pose significant risks to students' critical thinking and independent problem-solving abilities. In the context of STEM education, where multi-step reasoning and 'productive struggle' (Bjork & Bjork, 2011) are essential, this reliance creates a 'fragile knowledge' structure. Just as the Montessori method posits that intellectual independence is built through the 'mental hands-on' manipulation of concepts, AI-assisted homework risks bypassing this necessary cognitive labor, effectively replacing a scaffold (Wood et al., 1976) with a permanent cognitive prosthetic.

Engagement, Motivation, and the Illusion of Effectiveness

The reported increases in student engagement and the perceived usefulness of AI-assisted homework (Tamimi et al., 2024; Ward et al., 2025) present a psychological paradox. While AI reduces frustration by lowering the entry barrier to complex tasks, this heightened satisfaction often stems from the reduction of cognitive effort rather than a sense of mastery. As fundamentally argued by Soderstrom and Bjork (2015), high performance and positive affect during a task can lead to an "illusion of competence," where students mistake the ease of AI-facilitated task completion for genuine understanding. This "illusion of effectiveness" masks shallow processing with polished outputs, creating a facade of academic achievement that lacks conceptual depth. Furthermore, it is critical to recognize that motivation is not a constant variable. While students initially show high participation rates, this is often driven by the "novelty effect" of generative AI, which tends to diminish over time as the tool becomes a mundane utility (Bond et al., 2024). In contrast, creativity, critical thinking, and science process skills are the durable cornerstones for future problem-solving. If AI tools are used to automate the "messy" and non-linear inquiry process, the student's epistemic agency—their authority and responsibility in knowledge construction—is severely compromised.

For STEM education, which emerged as a response to the need for interdisciplinary synthesis, the risk of "cognitive offloading" is particularly high. To prevent this, AI use must be reframed through the lens of Distributed Cognition (Salomon, 1993). In this model, the AI is not a surrogate that provides answers, but a partner in a "thinking system" where the human remains the primary agent. Just as the Montessori approach advocates for the development of foundational skills through active manipulation, modern homework must require students to critique, justify, and extend AI outputs. This ensures that the technology supports the expansion of human intelligence rather than its contraction, preserving the creative struggle necessary for genuine scientific discovery.

Conditional Effects on Critical Thinking and Problem Solving

The impact of AI-assisted homework on higher-order skills, such as critical thinking and scientific reasoning, emerged as deeply conditional. Our findings suggest that AI tools do not inherently enhance or undermine cognitive development; rather, their effects are pedagogically mediated by instructional design and assessment alignment. Studies explicitly targeting problem-solving reported positive outcomes only when AI use was constrained by reflective activities and metacognitive prompts (Molenaar, 2022; Zhou et al., 2024). In contrast, overreliance on generative AI has been linked to reduced independent problem-solving and less diverse student responses, reflecting concerns that AI may supplant deeper cognitive engagement rather than support it (Qian, 2025). This underscores a critical "mediation effect": AI supports learning when it functions as a Socratic interlocutor that prompts the student to explain their reasoning, but it hinders learning when it serves as a mere content generator that provides final answers. This distinction is particularly vital for STEM education and the development of SPS. STEM inquiry is fundamentally interdisciplinary and non-linear, requiring students to engage in "messy" problem-solving that cannot be reduced to an algorithmic output. When assessments privilege surface features or final numerical answers, AI tools inadvertently discourage deeper engagement, leading to what Biggs (2003) termed "surface approaches to learning." However, when homework is designed to reward process explanation, justification, and conceptual transfer, AI can become a powerful partner in distributed cognition (Salomon, 1993). In this context, the student retains their epistemic agency by using AI to test hypotheses or

critique scientific models, rather than allowing the AI to bypass the "mental hands-on" struggle deemed essential by the Montessori philosophy.

Ultimately, the results challenge the current state of homework design. The perceived academic performance gains identified in this synthesis often reflect the AI's efficiency rather than the student's growth. To move beyond this "illusion of effectiveness," educators must align AI integration with assessment practices that value the learning process over the product. As argued by Lodge et al. (2023), the era of AI necessitates an assessment reform where students are evaluated on their ability to monitor, critique, and authorize knowledge. By forcing students to engage in higher-order epistemic monitoring, we ensure that creativity and critical thinking remain the cornerstones of scientific inquiry, preparing learners for the complex, interdisciplinary challenges of the future.

Implications for Interpreting Academic Performance Gains

Taken together, the results challenge simplistic interpretations of improved academic performance in AI-enhanced homework contexts. Performance gains reflected in grades, GPA, and completion metrics should be interpreted as indicators of enhanced productivity and output quality rather than unequivocal evidence of learning. This distinction is critical for educators and policymakers who may be tempted to equate measurable gains with educational success. The conditional patterns identified in this synthesis underscore the need to align AI integration with assessment practices that value learning processes, metacognitive regulation, and higher-order thinking. Without such alignment, AI homework tools risk reinforcing surface learning while inflating conventional performance metrics.

Implications

For educators, the findings underscore the importance of integrating AI homework tools within pedagogical frameworks that emphasize explanation, reflection, and revision. AI should be positioned as a support for learning rather than a substitute for effort. For researchers, future studies should prioritize delayed learning measures, examine interactions between AI use and metacognitive skills, and report effect sizes transparently. Policymakers should support human-centered AI integration, educator training, and long-term evaluation rather than relying on short-term performance metrics.

Limitations

This synthesis is limited by heterogeneity across included studies, incomplete reporting of effect sizes, and potential publication bias. Rapid technological change also limits the generalizability of findings to future AI systems with different capabilities.

A particularly significant limitation of this review is the pronounced imbalance in the disciplinary distribution of the evidence base. The synthesis provides robust empirical grounding for language education contexts—specifically EFL writing, academic writing in a second language, and translation tasks—where seven of the ten included studies were concentrated. In contrast, empirical research examining AI-assisted homework in STEM and pure science domains remains critically scarce. Only two studies (Kelly et al., 2013; AlShibli et al., 2025) explicitly addressed mathematics or computer science contexts, and neither reported sufficient effect size data to support strong domain-specific conclusions. No included studies examined AI homework tools in physics, chemistry, biology, or integrated STEM curricula. This disciplinary scarcity constitutes a formal limitation of the present study and must be clearly acknowledged when interpreting the generalizability of the findings.

The consequences of this "disciplinary divide" for the generalizability of the findings are substantial. The performance advantages associated with AI-assisted homework—particularly the large effect sizes documented in writing proficiency and language use—are closely tied to the surface-level features that generative AI tools are most adept at improving (i.e., grammatical accuracy, coherence, and organizational structure). These features are central to assessment in language education but are considerably less central to evaluation in STEM disciplines, where higher-order outcomes such as multi-step reasoning, mathematical proof, experimental design, and conceptual transfer are prioritized. Consequently, the predominantly positive findings of this synthesis should not be extrapolated to STEM educational contexts without direct empirical support. Future systematic reviews and primary studies should specifically target STEM domains to establish whether AI homework tools offer comparable, diminished, or qualitatively different benefits in contexts that privilege deep conceptual engagement

over linguistic output quality. Until such evidence is available, claims regarding the effectiveness of AI homework tools must be understood as domain-conditional, grounded predominantly in language and writing education rather than representing a universal finding across disciplines.

Conclusion

This study set out to examine the effects of AI-assisted homework on students' academic performance by synthesizing empirical evidence comparing AI-supported and traditional homework practices. Taken together, the findings indicate that AI tools are consistently associated with improvements in measurable academic outputs—such as assignment grades, task completion rates, and short-term assessment scores—but that these gains should be interpreted with caution when used as proxies for learning. A central conclusion emerging from this synthesis is that improvements attributed to AI-assisted homework are highly contingent on how academic performance is operationalized. Metrics commonly used in the reviewed studies tend to privilege efficiency, surface-level correctness, and linguistic or structural quality, areas in which generative AI systems are particularly effective. As a result, observed performance advantages often reflect enhanced productivity and output optimization rather than unequivocal gains in conceptual understanding, transfer, or long-term retention.

The analysis further demonstrates that the educational value of AI-assisted homework is not inherent to the technology itself but is mediated by pedagogical design, assessment alignment, and students' self-regulatory capacities. When AI use is embedded within instructional frameworks that emphasize reflection, justification, and process-oriented assessment, AI tools can function as cognitive scaffolds that support higher-order thinking. In contrast, unstructured or unrestricted AI use risks fostering cognitive offloading, homogenization of student work, and inflated performance indicators disconnected from deep learning. These findings carry important implications for both research and practice. For researchers, the results underscore the need to move beyond binary comparisons of AI versus non-AI conditions and toward more nuanced investigations that examine conditional effects, mediating variables, and the alignment between learning objectives and assessment practices. Future studies would benefit from incorporating longitudinal designs, process-based measures of learning, and explicit operationalizations of metacognitive engagement.

For educators and policymakers, the results caution against equating improvements in grades or completion rates with educational success. While AI-assisted homework can enhance accessibility, efficiency, and student engagement, its integration should be guided by clear pedagogical intentions and supported by assessment practices that value reasoning, originality, and epistemic agency. Without such safeguards, AI risks reinforcing surface learning while obscuring meaningful differences in students' understanding. In conclusion, AI-assisted homework represents neither a panacea nor a threat to learning outcomes. Its impact depends fundamentally on how it is designed, regulated, and evaluated within educational systems. By reframing academic performance as a multidimensional construct that extends beyond easily measurable outputs, this study contributes to a more balanced and theoretically grounded understanding of AI's role in contemporary education.

Scientific Ethics Declaration

* The authors declare that the scientific, ethical, and legal responsibility of this article published in the JESEH journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* There is no funding

References

AlShibli, A. S., Al Shibli, M., & Al Harthi, A. (2025). Exploring the impact of artificial intelligence on student academic performance. *Artificial Intelligence & Robotics Development Journal, 7*(1), 45–62. https://doi.org/10.52098/airdj.20233343

Biggs, J. (2003). *Teaching for quality learning at university* (2nd ed.). Society for Research into Higher Education & Open University Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *FABBS Foundation, Psychology and the real world: essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bognár, L., & Khine, M. S. (2025). The shifting landscape of student engagement: A pre–post semester analysis in AI-enhanced classrooms. *Computers and Education: Artificial Intelligence, 8*, 100395. https://doi.org/10.1016/j.caeai.2025.100395

Bond, M., et al. (2024). A systematic review of generative AI in higher education: The gap between opportunities and practice. *Applied Learning Review, 1*(1), 1-25. https://doi.org/10.1186/s41239-023-00436-z

Boumediene, H., & Bouakkaz, M. (2024). Changes in homework submission patterns with the advent of AI tools: A high school perspective. *Studies in Education Sciences, 5*(4), 112–129. https://doi.org/10.54019/sesv5n4-001

Chen, C., & Gong, Y. (2025). The role of AI-assisted learning in academic writing: A mixed-methods study on Chinese as a second language students. *Education Sciences, 15*(2), 141. https://doi.org/10.3390/educsci15020141

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research*, *76*(1), 1-62.

Dede, C. (2014). The role of digital technologies in deeper learning. Students at the Center: Deeper Learning Research Series. *Jobs for the Future*. 1-36.

Dettmers, S., Trautwein, U., & Lüdtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement*, *20*(4), 375-405. https://doi.org/10.1080/09243450902904601

Faridoon, N., Talpur, Q., Latif, F., Naz, G., & Shahzad, T. (2025). The role of AI tutors in improving academic performance and student engagement. *Academia International Journal for Social Sciences, 4*(3), 5897–5910. https://doi.org/10.63056/ACAD.004.03.0837

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). SAGE.

Gowtham, R., Iyer, S., & Krishnan, P. (2026). Measuring what matters: Academic performance, productivity, and learning in AI-enhanced education. *Computers & Education, 201*, 104789.

Hattie, J. (2009). The Black Box of Tertiary Assessment: An Impending Revolution. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P.M. Johnston, & M. Rees (Eds.), *Tertiary Assessment & Higher Education Student Outcomes: Policy, Practice & Research* (pp.259-275). Wellington, New Zealand: Ako Aotearoa

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, J., & Soffer Goldstein, D. (2013, July). Estimating the effect of web-based homework. *In International Conference on Artificial Intelligence in Education* (pp. 824-827). Berlin, Heidelberg: Springer Berlin Heidelberg. Aied2013ws_volume8.pdf

Khaustov, O. N., Tormyshova, T. Y., & Sukhanova, N. I. (2024). Teaching students to write academic papers through the use of generative artificial intelligence tools. Tambov University Review. *Series: Humanities, 29*(5), 1194-1207. https://doi.org/10.20310/1810-0201-2024-29-5-1194-1207

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A Meta-analytic review. *Review of Educational Research, 86*(1), 42-78. https://doi.org/10.3102/0034654315581420

Kwak, M. (2025). The effectiveness of AI-driven tools in improving student learning outcomes compared to traditional methods. *Issues in Information Systems, 26*(1), 1–12. https://doi.org/10.48009/4_iis_2025_120

Lodge, J., Howard, S., Bearman, M., & Dawson, P. (2023). *Assessment reform for the age of Artificial Intelligence.* Tertiary Education Quality and Standards Agency.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed. An argument for AI in education.* London: Pearson.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology, 46*(1), 4–11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632-645. https://doi.org/10.1111/ejed.12527

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*(n71). https://doi.org/10.1136/bmj.n71

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Publishing. https://doi.org/10.1002/9780470754887

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC methods programme*. University of Lancaster. https://doi.org/10.13140/2.1.1018.4643

Qian, Y. (2025). *Pedagogical applications of generative AI in higher education: A systematic review of the field*. TechTrends, 69, 1105–1120. https://doi.org/10.1007/s11528-025-01100-1

Salomon, G. (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge University Press.

Skinner, B. F. (1961). The Science of Learning and the Art of Teaching. In B. F. Skinner, *Cumulative record* (Enlarged ed., pp. 145–157). Appleton-Century-Crofts. https://doi.org/10.1037/11324-010

Smerdon, D. (2024). AI in essay-based assessment: Student adoption, usage, and performance. *Computers and Education: Artificial Intelligence, 5*, 100288. https://doi.org/10.1016/j.caeai.2024.100288

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Tamimi, J., Addichane, E., & Alaoui, S. M. (2024). Evaluating the effects of artificial intelligence homework assistance tools on high school students' academic performance and personal development. *Arab World English Journal, 15*(3), 45–67.

Trautwein, U., & Köller, O. (2003). The relationship between homework and achievement—still much of a mystery. *Educational Psychology Review*, *15*(2), 115-145.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Wale, B. D. (2024). Artificial intelligence in education: Effects of using integrative automated writing evaluation programs on honing academic writing instruction. *Cakrawala Pendidikan, 43*(1), 273-287.

Ward, B., Bhati, D., Neha, F., & Guercio, A. (2025). Analyzing the impact of AI tools on student study habits and academic performance. In *Proceedings of the IEEE 15th Annual Computing and Communication Workshop and Conference* (pp. 1–8).

Wood, D., 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

Yang, H. (2025). Harnessing generative AI: Exploring its impact on cognitive engagement, emotional engagement, learning retention, reward sensitivity, and motivation through reinforcement theory. *Learning and Motivation*, *90*, 102136. https://doi.org/10.1016/j.lmot.2025.102136

Yavich, R. (2025). Will the Use of AI undermine students independent thinking? *Education Sciences*, *15*(6), 669. https://doi.org/10.3390/educsci15060669

Zhou, X., Teng, D., & Al-Samarraie, H. (2024). The mediating role of generative AI self-regulation on students' critical thinking and problem-solving. *Education Sciences, 14*(9), 987.

Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students*, *29*(3), 42-46. https://doi.org/10.1145/3589649

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

Author(s) Information

Seyma Irmak
Amasya University,
Faculty of Education, Department of Mathematics and Science Education, Amasya/Türkiye
Contact e-mail: *seyma.bardak@gmail.com*
ORCID iD: https://orcid.org/0000-0003-3831-8244

Kaan Bati
Hacettepe University,
Faculty of Education, Department of Mathematics and Science Education, Ankara/Türkiye
ORCID iD: https://orcid.org/0000-0002-6169-7871

NOTE: Appendix added.

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	5-6-7
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	7-8
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	9-10
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	11-12-13
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	9-10-11
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	11-12-13
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	10-11-12-13
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	10-11-12
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	10-11-12-13
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	10-11-12
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	13
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	10-11-12-13
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	12-13
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	10-11-12
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to	14-15-16
		identify the presence and extent of statistical heterogeneity, and software package(s) used.	
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	16-24
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	16, 17, 18, 19, 20, 21, 22, 23 and 24
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	10-13
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	10-13
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	13-14
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	14
Study characteristics	17	Cite each included study and present its characteristics.	15
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	14-15
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	15-26
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	15-26
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	15-26
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	15-26
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	15-26
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	15-26
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	15-26
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	26-27
	23b	Discuss any limitations of the evidence included in the review.	32
	23c	Discuss any limitations of the review processes used.	32
	23d	Discuss implications of the results for practice, policy, and future research.	27-31
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	32
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	32
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	32
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	32
Competing interests	26	Declare any competing interests of review authors.	32
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	32