MISOOL: Jurnal Pendidikan Dasar
ISSN 2964-691X
Volume 8 Issue 2, 2026 https://doi.org/10.47945/misool.v8i2.3362

Artificial Intelligence in Science Lesson Planning: A Systematic Literature Review

Rabiudin[a] | **Mohammad Rian Ramli**[b]
[a]Science Education, Institut Agama Islam Negeri Sorong, Indonesia
[b]Primary Teacher Education, Institut Agama Islam Negeri Sorong, Indonesia
Correspondence: rabiudin27@gmail.com

Article History:
Received: 12/09/2026
Accepted: 26/11/2026

Keywords:
Artificial intelligence, ChatGPT, Lesson planning, Science teacher education, Systematic literature review

Abstract
Generative artificial intelligence (AI) has rapidly entered lesson planning, one of the most time-consuming aspects of teachers' work, yet evidence on how it shapes science lesson design remains fragmented. This systematic literature review synthesizes empirical studies published between 2020 and 2026 on the use of AI in planning K-12 science instruction. Following PRISMA 2020, 160 records were retrieved from Scopus and Google Scholar via Publish or Perish; after duplicate removal and two-stage screening by two reviewers ($\kappa = 0.87$), 18 studies from 11 countries were included and analyzed thematically. AI consistently accelerated drafting and broadened instructional ideas, but plan quality depended on prompt structure and the planner's pedagogical content knowledge. AI-generated plans were generally accurate in content yet weak in inquiry orientation, adaptivity, and attention to students' conceptions, and experienced teachers used AI suggestions more critically than novices. Training improved teachers' confidence and AI readiness, whereas gains in integrated professional knowledge were mixed. Five studies were Indonesian, but none examined teachers' actual use of generative AI in designing science teaching modules. The findings position AI as a co-designer whose contributions must be governed by teachers' pedagogical reasoning and call for teacher education that cultivates the critical evaluation of AI-generated plans.

© 2026 The Authors. This open access article is distributed under a (CC-BY License)

How to Cite:
Rabiudin, R., & Ramli, M. R. (2026). Artificial Intelligence in Science Lesson Planning: A Systematic Literature Review. MISOOL: Jurnal Pendidikan Dasar, 8(2), 289–300. https://doi.org/10.47945/misool.v8i2.3362

INTRODUCTION

Lesson planning lies at the heart of teachers' professional work, for it is at this stage that curriculum standards, disciplinary content, learners' prior conceptions, and assessment intentions are translated into a coherent sequence of classroom activity. In science education, this translation is especially demanding. A well-designed science lesson must represent natural phenomena accurately, anticipate alternative conceptions, structure inquiry or modelling activities, and align evidence of learning with the intended outcomes. It is therefore unsurprising that planning is widely regarded as one of the most time-consuming components of teaching. Powell and Courchesne (2024) describe lesson planning as essential yet labour-intensive, and they frame the search for efficiency without loss of quality as the main reason for examining generative artificial intelligence (AI) in this domain. When planning is rushed or reduced to generic templates, the consequences extend beyond administration: lessons may lose their conceptual precision and their capacity to engage students in authentic scientific practices. Any technology that promises to lighten the burden of planning should therefore be judged not only by the time it saves but also by whether it preserves or erodes the pedagogical reasoning that distinguishes a coherent science lesson from a mere list of activities.

Although AI in education is not new, its relationship with teachers has changed considerably. Earlier syntheses showed that most educational AI applications were designed for students, in the form of intelligent tutoring, adaptive systems, and automated assessment, while teachers were rarely positioned as primary users or designers (Zawacki-Richter et al., 2019). A subsequent review focusing on teachers identified potential benefits such as reduced workload and better-informed instructional decisions, but it also noted teachers' limited involvement in tool development and persistent concerns about the reliability of system outputs (Celik et al., 2022). The public release of large language models (LLMs) in late 2022 altered this landscape.

Conversational systems such as ChatGPT can generate learning objectives, activity sequences, worksheets, and assessment items on request, placing AI directly within the planning process rather than at its margins. Teachers have consequently become a central population for AI research, and new questions have emerged about professional knowledge, judgement, and responsibility.

The early scholarly response to generative AI combined enthusiasm with caution. Kasneci et al. (2023) argued that LLMs could help educators create materials and differentiate instruction, while warning that productive use depends on users' competencies and on their awareness of limitations such as bias and factual error. A rapid review by Lo (2023) likewise found that ChatGPT's performance varied across subject domains and that inaccurate or fabricated information was a recurring problem. Within science education, Cooper (2023) examined how ChatGPT responded to prompts about science teaching, including requests for units of work and assessment rubrics, and concluded that its outputs could serve as starting points but required critical evaluation by teachers. These contributions set the agenda, yet they were largely conceptual or researcher-driven. They showed what the tool could produce in principle without examining how teachers actually plan with it, what knowledge they bring to the interaction, or how the resulting plans measure up against disciplinary standards for science teaching.

Empirical research on AI-assisted science lesson planning has since accumulated, but it remains fragmented. Studies differ in population, ranging from pre-service to experienced teachers; in tools, from ChatGPT and Gemini to earlier intelligent tutoring systems; in science domains; and in design, from single cases to pre-post surveys. Their findings also diverge: some report high-quality plans and improved self-efficacy after structured exposure to generative AI, whereas others document fabricated resources, outdated curriculum references, and activities lacking genuine inquiry. Existing reviews only partly resolve this fragmentation. Systematic reviews published in Indonesian science education journals mapped the advantages and disadvantages of ChatGPT in science learning (Purnama et al., 2023; Asmiddin et al., 2023), but they were conducted when empirical studies were scarce, combined student-facing and teacher-facing uses, and did not treat lesson planning as a distinct professional task. Broader reviews of AI in education, in turn, seldom disaggregate findings by subject, even though the knowledge required to plan a science lesson differs substantively from that required in other disciplines.

Teacher knowledge offers a productive lens for interpreting this evidence. Pedagogical content knowledge (PCK) describes the knowledge that enables teachers to transform subject matter into forms that are teachable for particular learners, and the technological pedagogical content knowledge (TPACK) framework extends it to technology-rich teaching. Mishra et al. (2023) argued that generative AI shares the protean, opaque, and unstable qualities of earlier digital tools but differs in that it generates new content and invites conversational interaction, thereby altering the knowledge domains that teachers must coordinate. Celik (2023) proposed an Intelligent-TPACK model that adds ethical evaluation to the knowledge required for integrating AI-based tools, while Feldman-Maggor et al. (2025) showed, with chemistry examples, that responsible use also demands prompting skills and awareness of hallucination and bias. Complementing these accounts, Ng et al. (2023) described the AI digital competencies teachers require, and Zhai (2024) distinguished four roles in teachers' engagement with generative AI: observer, adopter, collaborator, and innovator. Together, these perspectives yield a testable expectation: if AI-assisted planning depends on teacher knowledge, the quality of AI-supported science lesson plans should vary with planners' content, pedagogical, and AI-related expertise.

The Indonesian context lends this question particular urgency. Under the current national curriculum, Indonesian teachers are expected to design their own teaching modules (*modul ajar*), a responsibility that increases planning workload and makes AI tools attractive. Indonesian research on AI-assisted science planning, however, predates generative AI: a series of studies documented a competency enhancement program in which West Java science teachers designed higher-order thinking skills (HOTS)-based integrated science lessons with the support of an intelligent tutoring system and web-based AI (Gunawan et al., 2020; Gunawan et al., 2021a). More recent Indonesian studies have examined what generative AI produces when asked to design ethnoscience or STEM lessons, but without observing teachers planning with it (Brar & Setioko, 2024; Utami et al., 2025). Much of the remaining Indonesian work on AI

and teaching modules appears in community service reports or Indonesian-language outlets. Indonesian teachers' actual use of generative AI in writing science modules thus remains largely undocumented in peer-reviewed international literature.

Three further gaps motivate this review. Methodologically, many studies rely on small samples, self-reported perceptions, and single-group designs, which limits inference about the effect of AI on planning quality. Conceptually, studies draw on different frameworks, including TPACK, PCK, the 5E instructional model, Universal Design for Learning, and the Technology Acceptance Model, without a shared account of what constitutes quality in an AI-assisted science lesson plan. Empirically, findings concerning teacher experience, the accuracy of AI-generated content, and the value of structured training have not been compared across studies. Accordingly, this review addresses four research questions. RQ1: What are the characteristics of empirical studies on AI in science lesson planning published between 2020 and 2026? RQ2: How do science teachers use AI across the stages of lesson planning? RQ3: What affordances and limitations of AI-assisted science lesson planning are reported? RQ4: What teacher-related and instructional conditions mediate the quality of AI-assisted science lesson plans?

In answering these questions, the review makes three contributions. Empirically, it offers an integrated account of a rapidly growing yet fragmented body of research drawn from Asia, Europe, Africa, and the Americas. Conceptually, it interprets the evidence through PCK and TPACK to clarify how teacher knowledge governs the contributions of AI to lesson design, providing a basis for a more precise definition of AI-assisted planning quality in science. Contextually, it characterizes the Indonesian evidence base, which is present but concentrated in pre-generative interventions and output analyses, and identifies research priorities that would allow Indonesian science teacher education to build on locally grounded findings rather than extrapolate from other systems. The remainder of the article describes the review protocol, presents the results for each research question, discusses their theoretical and practical implications, and concludes with directions for future research.

METHOD

This study is a systematic literature review reported in accordance with the PRISMA 2020 statement (Page et al., 2021). The procedure followed the stages commonly recommended for systematic reviews in the social sciences, namely formulating research questions, developing a search strategy, applying explicit eligibility criteria, appraising study quality, extracting data, and synthesizing findings (Xiao & Watson, 2019). Because the included studies were methodologically heterogeneous, encompassing qualitative case studies, content analyses of AI-generated documents, survey studies, and mixed-methods designs, a statistical meta-analysis was neither feasible nor appropriate. A narrative and thematic synthesis was therefore adopted, consistent with the aim of explaining how and under what conditions AI shapes science lesson planning rather than estimating a pooled effect. The four research questions guided every subsequent decision, from the selection of search terms to the categories used in data extraction.

Search strategy

The search was conducted on 25 November 2026 using Publish or Perish, which provides structured access to Scopus and Google Scholar and allows results to be exported in RIS format (Harzing & Alakangas, 2016). Scopus served as the principal source because of its controlled indexing and reproducible Boolean retrieval, whereas Google Scholar was used as a supplementary source for journals with less complete coverage, in line with evidence that Google Scholar is unsuitable as a principal search system for systematic reviews (Gusenbauer & Haddaway, 2020). The publication window was set from 2020 to 2026, with a maximum of 1,000 records per query. The search string combined four concept blocks: ("artificial intelligence" OR "generative AI" OR "ChatGPT") AND ("lesson plan" OR "instructional design" OR "teaching module") AND ("science education" OR "science teaching") AND "Indonesia". The query returned 60 records from Scopus and 100 from Google Scholar, yielding 160 records in total.

Eligibility criteria

Eligibility criteria were defined a priori across six dimensions (Table 1). The concept criterion

restricted the review to studies that explicitly examined the use, integration, or impact of AI in lesson planning, including instructional design, the preparation of teaching modules, and the design or evaluation of lesson plans before teaching. Studies of AI used by students, automated tutoring, AI-based learning media, or automated grading without teacher involvement in design were excluded. The context criterion required science content, and the population criterion required in-service or pre-service science teachers at primary or secondary level. Five studies in which researchers generated K-12 science lesson plans with AI, or in which experts evaluated AI-supported plans, were retained despite the absence of teacher participants because their unit of analysis was the science lesson plan itself. This decision is reported transparently so that readers can judge its influence on the synthesis.

Table 1. Eligibility criteria

Dimension	Inclusion	Exclusion
Concept	AI in lesson planning, instructional design, or module and assessment design before teaching	AI used only by students, automated tutoring, AI-based learning media, or automated grading without teacher design
Context	Science or science-focused STEM content	Non-science subjects or general education without a science context
Population	In-service or pre-service K-12 science teachers	Non-education university students or corporate training
Publication type	Peer-reviewed empirical journal articles and reputable proceedings with primary data	Books, chapters, editorials, theses, preprints, and review articles
Time span	2020–2026	Before 2020
Language	English	Other languages

Screening and inter-rater reliability

After six duplicate records were removed, five matched by title across databases and one duplicated within Google Scholar, screening proceeded in two stages. In the first stage, two reviewers independently examined the titles and abstracts of 154 records and excluded 125: 48 did not address AI in lesson planning, 34 concerned non-science subjects or higher-education teaching, 23 were reviews or bibliometric analyses, 15 were books, chapters, preprints, theses, or proceedings front matter, three lacked primary data, and two were citation-only records. In the second stage, 29 full texts were assessed and 11 were excluded: three on conceptual grounds, two on population grounds, three because the science context was absent or unconfirmed, one for language, and two because the full text or peer-review status could not be verified. Eighteen studies were ultimately included (Table 2). Agreement between reviewers on inclusion decisions, measured with Cohen's kappa, was $\kappa = 0.87$, which indicates strong agreement under the stricter benchmarks proposed by McHugh (2012).

Table 2. PRISMA 2020 flow of records

PRISMA stage	Records (n)
Identified through Scopus	60
Identified through Google Scholar	100
Duplicates removed	6
Screened (title and abstract)	154
Excluded at title and abstract	125
Full texts assessed for eligibility	29
Full texts excluded	11
Studies included in the synthesis	18

Quality appraisal, data extraction, and synthesis

The methodological quality of the included studies was appraised with the Mixed Methods Appraisal Tool (MMAT) version 2018, which provides five core criteria for each of five study designs (Hong et al., 2018). Studies were not excluded on the basis of quality; instead, appraisal results informed the weight given to each finding, so that claims resting mainly on self-reported or single-group data were stated cautiously. A

structured extraction form captured bibliographic details, country, science domain, participants, AI tool, theoretical framework, design, planning stage supported by AI, reported affordances and limitations, and moderating conditions. Extraction was guided by an analytical framework with three components derived from the literature reviewed above: the planning process, the teacher knowledge base that mediates AI use, and the quality of the resulting plan judged against science-specific criteria such as conceptual accuracy, curricular alignment, inquiry structure, and attention to alternative conceptions. Extracted data were synthesized through reflexive thematic analysis (Braun & Clarke, 2019), combining deductive codes from this framework with inductive codes emerging from the studies, and themes were compared across studies to identify convergent and divergent findings.

RESULTS AND DISCUSSION

Characteristics of the included studies (RQ1)

Eighteen studies met the eligibility criteria (Table 3), and their temporal distribution reveals two distinct phases. Three studies published in 2020 and 2021 examined intelligent tutoring systems and web-based AI within a single Indonesian teacher development program, whereas the remaining 15, published between 2024 and 2026, examined generative AI, with eight appearing in 2026 alone. The studies originated from 11 countries: Indonesia (five), Turkey (four), and one each from South Korea, Greece, Hong Kong (China), Germany, Malaysia, South Africa, the United States, India, and the Philippines. Seven studies involved pre-service teachers, five involved in-service teachers, one compared participants with different levels of experience, and five analyzed AI-generated or AI-supported plans without teacher participants. ChatGPT was explicitly named in 12 studies. In terms of design, six studies were qualitative, five combined qualitative and quantitative data, three were descriptive program evaluations, two were single-group pre-post studies, and two were descriptive analyses of AI outputs. No study used a randomized or controlled comparison between teachers planning with and without AI, and samples rarely exceeded 50 participants.

Table 3. Characteristics of the included studies (n = 18)

Study	Country	Participants / data	AI tool	Science focus	Design
Gunawan et al. (2020)	Indonesia	In-service teachers' lesson plans	Intelligent tutoring system	Integrated science	Descriptive analysis
Gunawan et al. (2021a)	Indonesia	29 in-service teachers	Web-based AI	Integrated science, HOTS	Descriptive program evaluation
Gunawan et al. (2021b)	Indonesia	In-service teachers (same program)	Web-based AI	Integrated science	Descriptive program evaluation
Lee & Zhai (2024)	South Korea	29 pre-service elementary teachers	ChatGPT	Multiple domains	Lesson plan analysis, TPACK rubric
Brar & Setioko (2024)	Indonesia	ChatGPT-generated lesson plans	ChatGPT	Ethnoscience	Qualitative exploratory
Peikos & Stavrou (2025)	Greece	Researcher-generated plans (Grade 5)	ChatGPT-4o	Floating and sinking	Qualitative content analysis
Aydin Gunbatar et al. (2025)	Turkey	6 participants, mixed experience	ChatGPT	Acids and bases	Multi-case study
Cheng (2025)	Hong Kong	4 in-service teachers, 30 Grade 7 students	ChatGPT	Density	Lesson study case study
Utami et al. (2025)	Indonesia	GPT-4o outputs (Year 8)	GPT-4o	Heat and temperature, STEM	Exploratory output analysis
Coşkun & Altan (2025)	Turkey	43 in-service teachers, 52 lesson plans	ChatGPT 4.5 (evaluator)	Engineering design in science	Mixed methods
Großmann et al. (2026)	Germany	60 AI-generated plans	ChatGPT-4o	Inquiry-based biology	Exploratory mixed methods
Chua et al. (2026)	Malaysia	14 pre-service chemistry teachers	ChatGPT	Chemistry	Qualitative questionnaire
Ndlovu et al. (2026)	South Africa	19 pre-service teachers	ChatGPT	Chemical equilibrium	Mixed methods, Rasch analysis
Allaire et al. (2026)	USA	Teacher candidates, three programs	ChatGPT	5E science lessons	Pre-post survey with open responses

Study	Country	Participants / data	AI tool	Science focus	Design
Goyal & Chander (2026)	India	25 pre-service teachers	ChatGPT, Gemini	Inclusive science, UDL	Mixed methods, pre-post
Ligsanan et al. (2026)	Philippines	6 pre-service teachers	AI tools	Physical and biological sciences	Descriptive phenomenology
Gürbüz et al. (2026)	Turkey	41 pre-service teachers	Generative AI tools	Curriculum-aligned science lessons	Single-group pre-post
Karaismailoglu et al. (2026)	Turkey	Expert evaluations of AI-supported plans	AI planning platforms	Engineering design-based learning	Descriptive expert evaluation

Uses of AI across planning stages (RQ2)

Across the studies, AI entered planning in three roles: as a generator of draft plans, as a co-designer in an iterative process, and, less commonly, as an evaluator of teacher-written plans. As a generator, AI produced complete plans from prompts, which researchers then appraised against disciplinary criteria (Peikos & Stavrou, 2025; Großmann et al., 2026; Utami et al., 2025). As a co-designer, AI was embedded in teachers' own design work. Lee and Zhai (2024) identified 14 types of teaching strategies in pre-service teachers' ChatGPT-integrated plans, which scored highest on the rubric dimension concerning instructional strategies but lower on exploiting the tool's full potential. In Hong Kong, four teachers used ChatGPT within a Lesson Study cycle, which reportedly made collaborative planning more efficient and enriched pedagogical discussion (Cheng, 2025). As an evaluator, ChatGPT 4.5 scored 52 engineering-integrated science lesson plans written by 43 teachers during professional development (Coşkun & Altan, 2025). In the Indonesian program, the AI system supported teachers' online learning and offered suggestions while they designed integrated science lessons (Gunawan et al., 2021a).

Reported affordances (RQ3)

The most consistently reported affordance was efficiency combined with a broader range of ideas. Malaysian pre-service chemistry teachers valued AI for structuring content, organizing instructional flow, generating activities, and saving time (Chua et al., 2026), and Filipino pre-service teachers considered it helpful for generating materials, aligning lessons with curricular standards, and supporting differentiated strategies (Ligsanan et al., 2026). A second affordance concerned plan quality under structured conditions. When pre-service teachers adapted ChatGPT output into their own plans, their reasoning improved in anticipating student thinking, justifying instructional decisions, and aligning tasks with learning goals (Ndlovu et al., 2026). PCK-informed, layered prompting with a reference paper improved alignment with research recommendations (Peikos & Stavrou, 2025), and structured prompts produced AI evaluations that agreed well with expert ratings (ICC = 0.708), whereas unstructured prompts did not (ICC = 0.076) (Coşkun & Altan, 2025). A third affordance was growth in confidence and readiness: teacher candidates reported greater AI familiarity, comfort, and perceived usefulness after a 5E planning activity (Allaire et al., 2026), and pre-service teachers reported an improved ability to design inclusive plans (Goyal & Chander, 2026).

Reported limitations clustered around inquiry quality, depth, and curricular fit. In the largest artifact analysis, ChatGPT-4o plans were generally accurate in content and aligned with learning outcomes, yet only 5 of 60 reached 75% of the maximum quality score and 22 fell below 50%, largely because the plans drifted from inquiry processes toward subject-matter acquisition and neglected students' inquiry-related conceptions (Großmann et al., 2026). Similar weaknesses appeared on a smaller scale: early outputs relied on a density formula poorly suited to primary pupils, assessments were mostly summative, and misconception-targeting activities appeared only after layered prompting (Peikos & Stavrou, 2025). Indonesian analyses reported that GPT-4o's pedagogical explanations lacked detail and citations (Utami et al., 2025) and that integrating local wisdom required deep cultural understanding and curricular alignment (Brar & Setioko, 2024). Accuracy problems also recurred, including AI-suggested alternative conceptions unsupported by the literature and formula errors (Aydin Gunbatar et al., 2025), fabricated internet material (Lee & Zhai, 2024), and factual inaccuracies, misalignment with the Malaysian curriculum, and impractical suggestions (Chua et al., 2026). AI evaluators, moreover, tended to be more positive and holistic than experts (Coşkun & Altan, 2025).

Conditions mediating planning quality (RQ4)

Three conditions recurred across the studies. The first was teaching experience. Aydin Gunbatar et al. (2025) identified four profiles, namely AI-reliant planners, AI-collaborative planners, AI-assisted plan refiners, and AI-independent planners, and found that more experienced participants filtered AI content more critically; only one participant systematically questioned its accuracy. The second was prompt structure combined with the planner's professional knowledge, evident in the gains from PCK-informed and structured prompting and in the conclusion of Großmann et al. (2026) that AI output requires substantial revision by teachers with strong PCK. The third was the type and duration of preparation. Interventions consistently raised confidence and readiness, but a 14-week program improved pre-service teachers' AI readiness without significantly changing their maker TPACK (Gürbüz et al., 2026), and adaptivity in AI-supported plans remained limited even when other reasoning dimensions improved (Ndlovu et al., 2026). Collaborative structures such as Lesson Study appeared to support productive use (Cheng, 2025).

AI as an amplifier of existing teacher knowledge

The central finding of this review is that AI did not equalize planning quality across teachers; rather, it tended to amplify differences in the knowledge that planners brought to the task. Experienced participants in the Turkish multi-case study filtered ChatGPT's content critically, whereas less experienced participants accepted most of it (Aydin Gunbatar et al., 2025). The German artifact analysis reached a convergent conclusion from a different direction: because ChatGPT-4o produced plans of highly variable quality, its output required careful evaluation and substantial revision by teachers with strong PCK (Großmann et al., 2026). Filipino pre-service teachers similarly emphasized the need for teacher judgement and noted the limited content depth of AI output (Ligsanan et al., 2026). This convergence supports the expectation derived from TPACK and its AI extensions, in particular the argument of Mishra et al. (2023) that the opacity and instability of generative AI place new demands on teachers' knowledge. One plausible interpretation is that AI shifts the cognitive work of planning from generation to evaluation, which is precisely the work that novices are least prepared to perform.

This interpretation carries an uncomfortable implication for pre-service teacher education. The group that stands to gain most from time savings, namely novices who find planning laborious, is also the group least able to detect what AI omits or distorts. The finding extends the argument of Kasneci et al. (2023) that productive use of LLMs depends on users' critical competencies by showing that this dependency is visible in authentic science planning tasks. The role framework of Zhai (2024) helps explain the profiles observed: AI-reliant planners resemble adopters who accept outputs, whereas plan refiners and AI-independent planners act as collaborators who retain authority over pedagogical decisions. The South African study suggests that this authority can be cultivated, since adapting ChatGPT output strengthened pre-service teachers' anticipation of student thinking and justification of decisions (Ndlovu et al., 2026). The pedagogical goal, therefore, is not to increase AI use as such but to move novice teachers toward collaborative forms of use in which AI output becomes an object of reasoning.

The tension between efficiency and disciplinary fidelity

A second theme concerns a tension between efficiency and disciplinary fidelity. Efficiency gains were reported almost universally, but they occurred at the level of drafting, whereas the features that distinguish strong science lessons, such as inquiry orientation, attention to students' conceptions, developmentally appropriate representations, adaptivity, and formative assessment, were those that AI handled least reliably. The clearest evidence comes from the 60 inquiry-based biology plans analyzed by Großmann et al. (2026), in which content accuracy was a strength but a consistent inquiry focus was not, and from the Greek study, in which misconception-targeting activities emerged only after layered prompting (Peikos & Stavrou, 2025). These results accord with Cooper (2023), who cautioned that ChatGPT's science teaching outputs require critical evaluation, and with Lo (2023), who documented variable performance and inaccurate information. A plausible explanation is that general-purpose LLMs reproduce the most common representations of school science in their training data, which tend toward textbook explanations rather than research-informed inquiry. Science-specific quality therefore does not emerge by default; it must be specified

by the planner.

The same tension appears when AI is used to evaluate rather than generate plans. AI evaluations agreed with expert ratings only under structured prompts and were more positive and holistic than expert feedback (Coşkun & Altan, 2025). Such leniency matters, because a teacher who relies on AI to check a plan may receive reassurance where an expert would identify weaknesses. The tension also helps explain why studies reached different conclusions about the value of AI. Studies relying on self-reported perceptions tended to report favourable outcomes (Allaire et al., 2026; Goyal & Chander, 2026), whereas studies evaluating plans against discipline-specific criteria more often identified substantive flaws. The divergence is thus partly methodological: what is found depends on what is measured. Claims about the benefits of AI for science planning should therefore be read alongside the criteria used to judge quality, and perception-based studies are likely to overestimate benefits relative to artifact-based evaluations.

Confidence is not competence

A third theme concerns the outcomes of training. Every intervention study reported gains in confidence, familiarity, or readiness, but evidence of gains in integrated professional knowledge was weaker. A 14-week program raised pre-service teachers' AI readiness without significantly changing their maker TPACK (Gürbüz et al., 2026), and adaptivity in lesson design remained limited even as other dimensions of pedagogical reasoning improved (Ndlovu et al., 2026). These findings suggest that becoming comfortable with AI and becoming able to integrate it into sound science pedagogy are distinct outcomes that develop at different rates. Readiness may be achieved through brief exposure, whereas integrated knowledge is likely to require sustained, practice-based experience with feedback. Collaborative structures may help in this respect: in the Hong Kong Lesson Study, ChatGPT supported planning within a cycle of joint design, implementation, and reflection, and the teachers reportedly grew in TPACK (Cheng, 2025). Because this evidence comes from a single case involving four teachers, it should be regarded as a promising direction rather than an established effect.

The Indonesian evidence base

The review identifies a contextual gap for Indonesia that is more specific than a simple absence of studies. Five of the 18 included studies were Indonesian, the largest national contribution, but three examined pre-generative AI within one teacher development program and may share participants (Gunawan et al., 2020; Gunawan et al., 2021a; Gunawan et al., 2021b), while the two generative-AI studies analyzed outputs without observing teachers (Brar & Setioko, 2024; Utami et al., 2025). No included study examined how Indonesian science teachers actually use generative AI to write *modul ajar*. Earlier Indonesian reviews (Purnama et al., 2023; Asmiddin et al., 2023) mapped general opportunities and risks but could not draw on such studies. This gap matters because the conditions identified here, particularly curricular fit and planner expertise, are context-dependent. The Indonesian ethnoscience study further showed that integrating local wisdom requires cultural understanding and curricular alignment that AI does not supply, a finding of direct relevance to the emphasis on contextual learning in Indonesian classrooms, including primary schools.

Theoretically, the review supports a view of AI-assisted science planning as distributed design in which the teacher's PCK and TPACK govern the machine's contributions. It also suggests that existing frameworks require a more explicit evaluative component. Feldman-Maggor et al. (2025) argued that prompting and awareness of hallucination fall partly outside TPACK; the present findings add the evaluation of inquiry orientation, curricular verification, and disciplinary error detection as competencies that AI-assisted planning makes newly central, and the competency framework of Ng et al. (2023) could be refined in this direction for science teachers. Practically, teacher education programs should treat AI not as a shortcut but as an object of critical study. Tasks in which pre-service teachers audit AI-generated plans against PCK criteria, inquiry rubrics, and curriculum documents could cultivate the evaluative stance that experienced teachers already display. This recommendation is consistent with the view of van den Berg and du Plessis (2023) that generative AI supports critical thinking in teacher education when its outputs are examined rather than simply adopted.

Several limitations qualify these conclusions. The evidence base is small, recent, and dominated by

single-site studies with modest samples, and no study compared teachers planning with and without AI under controlled conditions. The search relied on Scopus and Google Scholar through Publish or Perish; Google Scholar's limited Boolean functionality may have affected recall (Gusenbauer & Haddaway, 2020), and the search string did not retrieve several eligible studies known to the authors, such as a case study of ChatGPT-generated first-grade science plans (Powell & Courchesne, 2024) and a quasi-experiment on DeepSeek-assisted physics instructional design (Tong et al., 2026). Future updates should therefore incorporate citation searching and Web of Science. Restricting the review to English-language publications excluded Indonesian-language studies that may offer insight into local practice. Finally, retaining five studies without teacher participants broadened the evidence on plan quality, but their findings describe AI capabilities rather than teachers' typical use.

CONCLUSION

This review synthesized 18 empirical studies on the use of AI in science lesson planning published between 2020 and 2026. The evidence indicates that AI, particularly generative AI, reliably accelerates drafting and widens the range of instructional ideas available to science teachers, yet it does not by itself secure the features that define high-quality science lessons. Content accuracy was generally adequate, but inquiry orientation, attention to students' conceptions, adaptivity, and curricular fit depended on the knowledge and prompting of the planner, and experienced teachers filtered AI suggestions more critically than novices did. Training consistently improved confidence and AI readiness, while evidence of gains in integrated professional knowledge was mixed. The review thus contributes a knowledge-centred account of AI-assisted planning in which AI functions as a co-designer whose contributions must be governed by teachers' pedagogical reasoning rather than as a substitute for it. Indonesian studies formed the largest national subgroup, yet none examined teachers' actual use of generative AI in writing science modules. Teacher education programs should therefore engage pre-service teachers in critically auditing AI-generated plans, and future research should employ controlled designs, evaluate AI-assisted plans against discipline- and curriculum-specific criteria, and follow these plans into classroom enactment and student learning.

REFERENCES

Allaire, F. S., Pecore, J. L., & Sivaraj, R. (2026). From Skepticism to Innovation in 5E-Asy Steps: Harnessing ChatGPT for 5E Science Lesson Planning. Journal of College Science Teaching, 55(4), 378-386. https://doi.org/10.1080/0047231X.2026.2650312

Asmiddin, A. M., Nurhuda, P., & Megawati, R. (2023). Advantages and disadvantages of ChatGPT in science learning: A systematic literature review. *Jurnal Penelitian Pendidikan IPA, 9*(12), 1335–1341. https://doi.org/10.29303/jppipa.v9i12.6576

Aydin Gunbatar, S., Tezcan Sirin, G., Ilkyaz, O. C., & Mutlu, Y. (2025). Exploring the artificial intelligence interaction profiles of participants with different levels of teaching experience for lesson planning in the context of acids and bases. *Chemistry Education Research and Practice, 26*(4), 977–995. https://doi.org/10.1039/d5rp00064e

Brar, J. H. S., & Setioko, W. (2024). ChatGPT and local wisdom: An exploratory study of AI-assisted ethnoscience lesson plan. In *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 294–299). IEEE. https://doi.org/10.1109/iSemantic63362.2024.10762585

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior, 138*, Article 107468. https://doi.org/10.1016/j.chb.2022.107468

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*(4), 616–630.

https://doi.org/10.1007/s11528-022-00715-y

Cheng, E. C. K. (2025). Leveraging generative AI in science lesson study: Transforming density concept instruction through ChatGPT integration. *International Journal for Lesson and Learning Studies, 14*(3), 215–236. https://doi.org/10.1108/IJLLS-11-2024-0277

Chua, K. H., Sathasivam, R. V., Mafarja, N., & Abdul Rahim, S. S. (2026). Exploring the benefits and challenges of AI-driven lesson planning among preservice science teachers. *STEM Education, 6*(1), 1–20. https://doi.org/10.3934/steme.2026001

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*, 444–452. https://doi.org/10.1007/s10956-023-10039-y

Coşkun, T. K., & Altan, E. B. (2025). Comparative analysis of AI and expert evaluations in engineering design pedagogy. *PLOS ONE, 20*(9), Article e0332715. https://doi.org/10.1371/journal.pone.0332715

Feldman-Maggor, Y., Blonder, R., & Alexandron, G. (2025). Perspectives of generative AI in chemistry education within the TPACK framework. *Journal of Science Education and Technology, 34*(1), 1–12. https://doi.org/10.1007/s10956-024-10147-3

Goyal, Y., & Chander, S. (2026). GenAI-supported approach to lesson planning for inclusive science classrooms. *Asia Pacific Journal of Educators and Education, 41*(1), 369–393. https://doi.org/10.21315/apjee2026.41.1.16

Großmann, L., Koberstein-Schwarz, M., Krüger, D., & Krell, M. (2026). Lesson planning with ChatGPT for inquiry-based biology instruction – A(I) roll of the dice? *International Journal of Science Education, 48*(15), 2356–2375. https://doi.org/10.1080/09500693.2025.2567509

Gunawan, K. D. H., Liliasari, S., Kaniawati, I., & Setiawan, W. (2020). Exploring science teachers' lesson plans by the implementation of intelligent tutoring systems in blended learning environments. *Universal Journal of Educational Research, 8*(10), 4776–4783. https://doi.org/10.13189/ujer.2020.081049

Gunawan, K. D. H., Liliasari, L., Kaniawati, I., & Setiawan, W. (2021a). Implementation of competency enhancement program for science teachers assisted by artificial intelligence in designing HOTS-based integrated science learning. *Jurnal Penelitian dan Pembelajaran IPA, 7*(1), 55–65. https://doi.org/10.30870/jppi.v7i1.8655

Gunawan, K. D. H., Liliasari, Kaniawati, I., & Setiawan, W. (2021, November). The responses to artificial intelligence in teacher integrated science learning training program. In Journal of Physics: Conference Series (Vol. 2098, No. 1, p. 012034). IOP Publishing. https://doi.org/10.1088/1742-6596/2098/1/012034

Gürbüz, F., Bostan Sarıoğlan, A., & Akbay, B. (2026). The effect of the AI-supported lesson planning process on science teacher candidates' technological pedagogical content knowledge and AI readiness. *Journal of Pedagogical Sociology and Psychology, 8*(3), Article e47606. https://doi.org/10.33902/JPSP.202647606

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Karaismailoglu, F., Surmeli, H., & Yildirim, M. (2026). Suitability of Artificial Intelligence Supported Lesson Plans from the Perspective of Science Education Experts. *Journal of Science Education and Technology*, 1-13.. https://doi.org/10.1007/s10956-026-10342-4

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.

Lee, G. G., & Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. IEEE Transactions on Learning Technologies, 17, 1643-1660. https://doi.org/10.1109/TLT.2024.3401457

*Ligsanan, L. S., Batchar, R., & Manguil, C. J. (2026). Exploring the influence of artificial intelligence adaptation on pre-service science teachers' pedagogical content knowledge. *Suranaree Journal of Social Science, 20*(2), Article e282481. https://doi.org/10.55766/sjss282481

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and generative AI. *Journal of Digital Learning in Teacher Education, 39*(4), 235–251. https://doi.org/10.1080/21532974.2023.2247480

*Ndlovu, B. P., Khoza, H. C., & Sibanda, D. (2026). Pre-service science teachers' adaptive pedagogical reasoning in AI-supported lesson plans: A case with chemical equilibrium. *Frontiers in Education, 11*, Article 1811975. https://doi.org/10.3389/feduc.2026.1811975

Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational Technology Research and Development, 71*(1), 137–161. https://doi.org/10.1007/s11423-023-10203-6

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

*Peikos, G., & Stavrou, D. (2025). ChatGPT for science lesson planning: An exploratory study based on pedagogical content knowledge. *Education Sciences, 15*(3), Article 338. https://doi.org/10.3390/educsci15030338

Powell, W., & Courchesne, S. (2024). Opportunities and risks involved in using ChatGPT to create first grade science lesson plans. *PLOS ONE, 19*(6), Article e0305337. https://doi.org/10.1371/journal.pone.0305337

Purnama, I., Edi, F., Megahati S, R. R. P., & Wijanarko, T. (2023). ChatGPT for teachers and students in science learning: A systematic literature review. *Jurnal Penelitian Pendidikan IPA, 9*(10), 760–765. https://doi.org/10.29303/jppipa.v9i10.5259

Tong, D., Wu, H., Islam, A. Y. M. A., Liu, J., Jiang, L., & Ren, H. (2026). DeepSeek-assisted physics instructional design: An empirical study of high school physics teaching. *Physical Review Physics Education Research, 22*, Article 010151. https://doi.org/10.1103/bxql-gzfk

Utami, A., Dhitareka, P. H., Husna, H. N., & Prima, E. C. (2025). An exploratory study of ChatGPT in STEM teaching on heat and temperature topic. Discover Education, 4(1), 309.. https://doi.org/10.1007/s44217-025-00751-9

van den Berg, G., & du Plessis, E. (2023). ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences, 13*(10),

Article 998. https://doi.org/10.3390/educsci13100998

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research, 39*(1), 93–112. https://doi.org/10.1177/0739456X17723971

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X. (2024). Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-024-10174-0