Journal of Science Education and Technology
https://doi.org/10.1007/s10956-026-10342-4

Suitability of Artificial Intelligence Supported Lesson Plans from the Perspective of Science Education Experts

Fatma Karaismailoglu[1] · **Hikmet Surmeli**[1] · **Mehtap Yildirim**[2]

Received: 17 August 2025 / Accepted: 8 June 2026
© The Author(s) 2026

Abstract
The growing integration of Artificial Intelligence (AI) in education has brought increased interest in using AI-supported platforms for lesson planning. However, concerns remain regarding the pedagogical validity and theoretical alignment of AI-generated plans. This study examines the instructional quality of lesson plans created by two AI platforms—one general-purpose and one education-focused—through expert evaluation. Utilizing a case study design, lesson plans based on the Engineering Design-Based Learning (EDBL) approach were generated for a sixth-grade science unit, incorporating four skill dimensions from the revised Turkish science curriculum. Eleven science education experts assessed each plan using a standardized rating scale and open-ended feedback. Findings reveal that while both platforms support key competencies such as conceptual and social-emotional skills, they fall short in fully capturing the iterative and reflective processes of EDBL. The education-focused platform demonstrated higher alignment with curriculum goals and instructional sequencing. Experts emphasized that AI-generated plans can serve as useful drafts but require pedagogical adaptation and teacher judgment for effective classroom use. The study highlights the promise and limitations of AI tools in science instruction, offering insights for educators, AI developers, and curriculum designers aiming to integrate AI meaningfully into teaching practice.

Keywords Artificial Intelligence · Engineering Design-Based Learning Approach · Lesson Planning · Science Education · Skills

Introduction

The rapid emergence of Artificial Intelligence (AI) tools in education has led to a growing interest in understanding their potential to support instructional design and planning. Among the various applications of AI, the use of generative AI platforms to produce lesson plans has gained significant attention due to their efficiency, adaptability, and accessibility for teachers (Hashem et al., 2024; Rasul et al., 2023). These platforms offer automated instructional materials by processing user prompts, thereby providing teachers with ready-to-use or modifiable learning scenarios. However, questions remain regarding the pedagogical soundness, contextual appropriateness, and theoretical alignment of such AI-generated plans.

Existing research has primarily focused on the usability of AI tools in designing lessons, particularly in the context of pre-service teacher education (Lee et al., 2024; van den Berg & du Plessis, 2023). While such studies highlight the potential benefits of AI in saving time and enhancing creativity, they often overlook the extent to which these tools align with established learning theories and curriculum standards. Specifically, little attention has been paid to how AI-generated lesson plans accommodate complex, process-oriented instructional approaches, such as the Engineering Design-Based Learning (EDBL) model—a pedagogy that emphasizes iterative problem-solving, creativity, and contextualized application in science education (Ndihokubwayo et al., 2022).

In response to this gap, the present study investigates the suitability of AI-supported lesson plans from the perspective of science education experts, with a particular focus on

✉ Fatma Karaismailoglu
fkaraismailoglu@mersin.edu.tr

[1] Department of Science Education, Faculty of Education, Mersin University, Mersin, Turkey

[2] Department of Science Education, Faculty of Education, Marmara University, Istanbul, Turkey

Published online: 18 June 2026

their alignment with the EDBL approach and competency-based skill development. The study compares two different AI platforms—one general-purpose and the other education-focused—by examining how effectively each platform constructs lesson plans that address both the stages of EDBL and the multidimensional skills emphasized in the updated national science curriculum in Turkey.

This research seeks to address the following question:

> What is the pedagogical quality and EDBL alignment of AI-generated lesson plans for a 6th-grade science unit, as assessed against national curriculum skills and EDBL stages?

By combining expert evaluations with descriptive analysis, the study aims to provide both empirical and theoretical insights into the pedagogical utility and limitations of AI-generated lesson plans in science education. Ultimately, this work contributes to the emerging discourse on AI integration in curriculum design, offering implications for educators, curriculum developers, and AI tool designers.

Theoretical Framework

Engineering Design-Based Learning (EDBL)

Engineering Design-Based Learning (EDBL) is a pedagogical approach grounded in problem-based learning and constructivist theories, wherein students engage in authentic, iterative design challenges to develop scientific understanding and practical skills (English & King, 2015). The EDBL process typically involves six core stages: (1) identifying the problem, (2) researching the problem, (3) proposing possible solutions, (4) selecting and developing the best solution, (5) prototyping and testing, and (6) evaluating and redesigning (Ndihokubwayo et al., 2022). In science education, EDBL fosters systems thinking, collaboration, and creative problem-solving—skills deemed essential for STEM literacy in the 21 st century.

However, EDBL demands instructional materials that are context-sensitive, cognitively engaging, and pedagogically sequenced. The success of this approach heavily depends on well-designed lesson plans that not only guide students through each phase but also align with the specific cognitive and affective goals of the curriculum. The applicability of AI in this context raises critical questions about whether such platforms can meaningfully replicate or support this nuanced learning design.

AI-Supported Instructional Design

AI-generated instructional content, particularly lesson plans, operates through language models that predict content based on user prompts (Davis & Lee, 2023). While these tools show potential in scaffolding routine planning tasks, they are not inherently pedagogical entities. Their outputs reflect statistical patterns, not necessarily instructional logic. Consequently, the evaluation of AI-generated lesson plans must go beyond surface-level usability and examine their alignment with theoretical models of learning and teaching (Carr, 2023; Gupta et al., 2023).

Recent studies have raised concerns regarding the limitations of AI in customizing content to local contexts, supporting higher-order thinking, and integrating socio-emotional or domain-specific skills (Truong et al., 2023; Rasul et al., 2023). These shortcomings are particularly salient in constructivist approaches like EDBL, where learning is context-rich and learner-driven.

Competency-Based Science Curriculum in Turkey

Turkey's 2024 revised secondary science curriculum adopts a competency-based framework that emphasizes four skill domains: (1) conceptual skills (e.g., problem-solving, questioning), (2) domain-specific scientific practices (e.g., observation, inference), (3) literacy skills (e.g., data, visual, and digital literacy), and (4) social-emotional learning skills (e.g., empathy, communication). These skills are to be addressed in parallel with subject matter content through active, student-centered instructional designs.

Given this curricular emphasis, evaluating whether AI-generated lesson plans adequately support these skill domains is essential. It is especially important in science education, where skills and process knowledge are as critical as conceptual understanding. Thus, this study situates AI-supported instructional design within both an international pedagogical discourse and a national curriculum reform context.

Methodology

Research Design and Context

This study employed a qualitative case study design with descriptive quantitative elements (Yin, 2018) to evaluate the pedagogical appropriateness and skill alignment of lesson plans generated by two different AI platforms—one general-purpose (C1) and one education-focused (C2). The research adopted a comparative and interpretive approach, focusing on how experts assess AI-generated lesson plans within the framework of the Engineering Design-Based Learning (EDBL) model and the skills emphasized in the national science curriculum. Given the exploratory nature of the research and the small, purposive expert sample, the analysis primarily focused on qualitative insights, with descriptive statistics serving to illustrate trends and patterns rather than to establish statistical significance.

The context of the study is grounded in the 2024 draft Turkish secondary science curriculum, which introduces a skills-based instructional paradigm. The curriculum categorizes skills into four dimensions:

- Conceptual skills, including high-order thinking abilities such as questioning, discussing, and problem-solving;
- Domain-specific skills, such as hypothesis formulation, observation, and inference;
- Literacy skills, referring to data, media, and visual literacy in technology-rich contexts;
- Social-emotional learning skills, such as emotional regulation, empathy, and healthy self-concept development.

To align with this framework, both AI platforms were asked to generate a lesson plan for 6th-grade students covering the "Sustainable Living and Biodiversity" unit. The specific learning objective addressed was:

> Ability to question the importance of biodiversity for natural life.

The full prompt was standardized across platforms:

> Q1. Generate a lesson plan for 6th grade students that includes the learning goal of *"questioning the importance of biodiversity for natural life"*, addressing the following sub-goals:
>
> a) defines the importance of biodiversity for natural life;
> b) asks questions about biodiversity;
> c) collects relevant information;
> d) evaluates the reliability of collected information;
> e) makes inferences based on the collected information.
>
> Q2. The lesson should incorporate all four skill dimensions: conceptual, domain-specific, literacy, and social-emotional learning skills.
> Q3. Assume that the lesson will be implemented in Mersin, Turkey, and revise the plan to reflect this local context.

This structure was intended to test whether the AI platforms could customize outputs to cultural and regional contexts—a known limitation in AI-generated educational content.

The prompt was developed iteratively with input from three science education experts to ensure it captured all five sub-learning goals and the EDBL framework. Initial prompts were written in Turkish; however, pilot testing revealed that both AI platforms (ChatGPT-4 and Teacher's Buddy) failed to adequately interpret the EDBL framework when prompted in Turkish. The generated lesson plans did not reflect the iterative, process-oriented nature of EDBL. Switching to English prompts produced substantially more pedagogically coherent outputs aligned with EDBL stages. Consequently, the final prompt was written in English and designed to be highly structured, specifying the target grade level (6th grade), the science unit, and the required skill dimensions. The prompt development process involved two rounds of expert review to ensure clarity, completeness, and alignment with curriculum objectives. No AI platform was used to refine the prompt beyond expert judgment.

Sampling of AI Platforms

Platform C1 was ChatGPT-4 (OpenAI, GPT-4 version), a general-purpose AI chatbot. Platform C2 was Teacher's Buddy (teachersbuddy.com, Lesson Plan Generator tool), a dedicated educational AI tool designed specifically for lesson planning. Both tools have been previously identified in the literature as having potential in instructional design (Karaman & Göksu, 2024; Rasul et al., 2023).

Participant Sampling

Expert participants were selected through criterion sampling (Patton, 2015). Inclusion criteria required at least a master's degree in science education or curriculum studies and demonstrated experience with the engineering design-based learning model. Eleven experts (coded as E1–E11) were selected accordingly. Their institutional affiliations, academic titles, and research expertise are presented in Table 1.

Data Collection and Evaluation Framework

The AI-generated plans were coded as C1 (general-purpose) and C2 (education-focused). Each expert evaluated both plans using a standardized form.

Evaluation Framework: The assessment was designed to capture both theoretical alignment and practical utility.

EDBL Stage Alignment: The evaluation was primarily grounded in the six-stage EDBL model. However, for enhanced analytical precision, the closely related stages of 'Proposing the solutions' and 'Choosing the best solution' were combined into a single rating category: 'Proposing Solutions and Choosing the Best Solution.' This adaptation reflected the structure of the AI-generated outputs and improved rating consistency.

Implementation Quality: To assess the practical classroom readiness of the plans, two additional criteria were evaluated: 'Presentation of the Solution' (assessing the structure and mode of final output presentation) and 'Student Assessment Methods' (evaluating the embedded strategies for evaluating student learning).

Skill Domain Coverage: The rubric evaluated five skill domains: the four curriculum-mandated competencies plus

Table 1 Information about experts

Expert No	Title	Profession	AI -based platform usage status	Platforms used
E1	PhD	Sci. Edu.	Yes	ChatGPT
E2	Res. Asst.	Sci. Edu.	Yes	ChatGPT
E3	PhD	Sci. Edu.	Yes	ChatGPT, Coplot
E4	Assoc. Prof.	Sci. Edu.	Yes	ChatGPT, Autodraw, Playground etc.
E5	MA	Sci. Edu.	Yes	ChatGPT
E6	MA	Sci. Edu.	Yes	ChatGPT, Deepl Translate
E7	MA	Sci. Edu.	Yes	ChatGPT
E8	MA	Sci. Edu.	Yes	ChatGPT
E9	PhD	Sci. Edu.	Yes	ChatGPT, Bing Creator
E10	MA	Sci. Edu.	No	-
E11	MA	Sci. Edu.	Yes	ChatGPT, Midjourney, Scispace

'Engineering Skills,' included as an inherent requirement of the EDBL framework.

Therefore, the quantitative analysis reports on seven main items: five EDBL-based stages and two implementation criteria, across five skill domains.

Each expert was provided with both plans and a standardized evaluation form, consisting of:

The AI-generated lesson plans were produced in English, as the education-focused platform (Teacher's Buddy) did not support Turkish language input at the time of the study. The English outputs were then independently translated into Turkish by two of the authors, both fluent in English and experts in science education. The translations were compared for consistency, and any discrepancies were resolved through discussion. The final Turkish versions of the lesson plans were presented to the expert evaluators. We acknowledge that translation may introduce minor interpretive shifts; however, the translation process was documented and reviewed to minimize alterations to the original AI output. This limitation is addressed in Sect. 5.2.

Rating Scale: A 10-point rubric (1 = not appropriate at all, 10 = fully appropriate) assessing twelve instructional quality criteria across five skill domains. Table 2 provides a detailed description of each criterion, including anchor examples for a score of 10. For analytical reporting, these were grouped into seven main categories (five EDBL-based stages and two implementation criteria). To prevent familiarity bias, the two lesson plans were presented to experts as unlabeled documents (coded as "C1" and "C2") with no indication of which AI platform generated which plan. Experts were not informed about the source of either plan, either in writing or orally.

Open-Ended Prompts: Experts were asked to explain their ratings and indicate platform preference with justification.

Quantitative Data Analysis

Quantitative ratings were analyzed descriptively (mean) for each plan across two main dimensions: (1) the seven instructional quality criteria and (2) the five skill categories. Given the study's qualitative orientation and small expert sample, these descriptive statistics were used solely to identify patterns and trends rather than to establish inferential significance.

Qualitative Data Analysis

Open-ended responses were subjected to thematic content analysis using Braun and Clarke's (2006) framework. The analysis followed a systematic six-phase approach:

Data Familiarization Two researchers independently immersed themselves in the data by repeatedly reading all expert responses.

Initial Coding Researchers independently generated codes capturing key concepts using a line-by-line analysis approach. Representative codes included 'needs modification,' 'lacks originality,' and 'useful draft.'

Theme Development Through constant comparison and discussion, researchers grouped related codes into potential themes. The coding process was iterative and reflexive, with researchers maintaining analytical memos to document decision-making.

Theme Review Preliminary themes were critically examined against the dataset to ensure accurate representation. This involved checking for coherence, distinctiveness, and comprehensive coverage of the data.

Theme Definition and Refinement The three final themes were clearly defined and refined: 'Applicable' (direct classroom readiness), 'Can be applied by correction' (usability with modifications), and 'Should not be applied' (rejection of classroom use).

Reporting Findings were organized around these refined themes with representative expert quotations.

Inter-Coder Reliability and Consensus The evaluation of the 11 experts' feedback involved a qualitative coding process where two researchers independently categorized the responses into three ordinal levels: 'Applicable', 'Can be applied by correction', and 'Should not be applied'. To ensure the reliability of this coding process, quadratic weighted kappa was calculated. The two researchers achieved a weighted κ of 0.79 (95% CI: 0.68–0.90), indicating substantial agreement beyond chance. All initial coding

Table 2 Dimensions and descriptors of the expert evaluation rubric

Criterion Category	Criterion Description	Score (1–10)	Anchor Example (10 = Fully appropriate)
EDBL Stage Alignment	The extent to which each EDBL stage (Identifying the Problem, Researching, Proposing & Selecting Solutions, Prototyping & Testing, Evaluating & Retesting, Presenting the Solution) follows the engineering design-based learning model.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*The activity clearly guides students to define a local biodiversity problem (e.g., in Mersin) and explicitly follows the iterative EDBL cycle.*
Skill Domain Coverage	For each EDBL stage, the plan's adequacy in developing: - Conceptual skills - Domain-specific skills - Literacy skills - Social-emotional learning skills - Engineering skills	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Students work in groups to evaluate conflicting information about biodiversity, requiring both collaboration (SEL) and critical evaluation (literacy).*
Implementation Quality – Presentation of Solution	Clarity, structure, and mode of final output presentation.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*The plan includes a well-defined format (e.g., infographic, poster, oral presentation) with clear criteria for effective communication.*
Implementation Quality – Student Assessment Methods	Embedded strategies for evaluating student learning (formative/summative).	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*The rubric includes peer feedback, self-reflection on prototype revisions, and alignment with learning goals.*
Contextual Customization	How well the plan adapts to the local context (Mersin, Turkey) in terms of examples, locations, and cultural relevance.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Activities reference local species, the school garden, or nearby parks in Mersin; examples are specific and authentic.*
Duration & Pacing	Appropriateness of the estimated time for each activity and the overall lesson.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*The allocated time (e.g., 105 min or 3–4 lessons) realistically allows for all EDBL stages without rushing.*
Material Availability	Whether required materials are accessible in a typical Turkish middle school.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Uses basic stationery, internet access, local guides – no specialized or expensive lab equipment needed.*
Attention Grabbing & Motivation	The plan's potential to engage students and sustain interest in STEM/biodiversity.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Opens with a local environmental issue (e.g., changes in Mersin's coastline) that students can personally relate to.*
Increasing Interest in STEM	The extent to which the plan encourages students' interest in science, technology, engineering, and mathematics.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Students act as engineers designing solutions for real biodiversity threats, fostering STEM career awareness.*
Value Acquisition	The plan's effectiveness in fostering environmental awareness, responsibility, and ethical understanding.	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Students conclude by proposing actionable, value-driven solutions for local biodiversity protection.*
Overall Learning Outcome Alignment	How well the plan addresses the specified learning goals (defining, questioning, collecting, evaluating, inferring about biodiversity).	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Each sub-goal is explicitly covered and assessed through EDBL stages, not merely listed.*
Grade Level Appropriateness	Suitability of content, language, and cognitive demand for 6th-grade students (age 11–12).	1 = Not appropriate at all 5–6 = Moderately appropriate 10 = Fully appropriate	*Vocabulary, activity complexity, and prior knowledge assumptions are appropriate for typical 6th graders.*

disagreements were resolved through discussion until full consensus was reached, ensuring the analytical credibility and trustworthiness of the findings.

Findings

This section presents the comparative evaluation of two AI-generated lesson plans—C1 (general-purpose) and C2 (education-specific)—across three dimensions: EDBL alignment, skill development adequacy, and contextual applicability. In line with the study's qualitative case study approach with descriptive quantitative elements, the findings focus on identifying patterns and trends in expert evaluations. Quantitative comparisons are presented as descriptive statistics to illustrate these patterns, while qualitative insights from open-ended responses provide depth and context to the numerical data.

Platform Comparison: EDBL Stage Alignment

Figure 1 presents the bubble chart of expert evaluations for C1, with the X-axis representing EDBL stages and the Y-axis denoting each expert. To aid interpretation, each bubble represents one expert's total score for a given stage—the larger the bubble, the higher the score (maximum 60 points per stage, 40 for "Evaluation").

Alt. Text. Bubble chart showing expert scores (E1–E11) for each EDBL stage in C1. Larger bubbles indicate higher

scores (max 60; ER max 40). Highest mean at Presenting Solution, lowest at Evaluating & Retesting.

Figure 1 shows individual expert ratings for C1 across EDBL stages, with bubble size representing score magnitude. Notably, the consistently large bubbles in the 'Presenting Solution (PS)' stage indicate strong and uniform expert agreement on C1's performance in final presentation. Conversely, the smaller and more varied bubble sizes in 'Evaluating & Retesting (ER)' reveal not only lower average scores but also greater disagreement among experts, highlighting C1's weakness in supporting iterative feedback processes. This pattern aligns with findings from van den Berg and du Plessis (2023) regarding AI limitations in facilitating reflection and revision.

Figure 2 visualizes expert ratings of C2 across the EDBL stages and implementation criteria.

To aid interpretation, each bubble represents one expert's total score for a given stage—the larger the bubble, the higher the perceived instructional quality for that stage (maximum 60 points per stage; 40 for "Evaluation & Retesting").

Bubble visualization was preferred over bar charts to simultaneously display inter-expert variation and overall stage trends.

Alt. Text: Bubble chart visualising the score given by each expert to C2 for each stage of the investigated learning approach.

The consistently large bubbles across all stages—particularly for *Presenting the Solution* ($\bar{x}$ = 49.09) and *Evaluating & Retesting* ($\bar{x}$ = 48.45)—demonstrate C2's robust and uniform performance throughout the learning cycle. Stronger bubbles in iterative phases like *Evaluating & Retesting* and *Making/Testing Prototype* ($\bar{x}$ = 47.45) visually confirm its stronger capacity for supporting reflective, process-oriented learning tasks. The education-focused architecture of C2 appears to provide more balanced coverage across all EDBL stages.

The comparison of C1 and C2 across EDBL stages (see Figs. 1 and 2) reveals distinct platform strengths. While both scored highest for "Presenting the solution," C1 was stronger in early stages such as "Identifying the problem," whereas C2 received higher scores in feedback-intensive stages like "Evaluating and retesting." These results reflect C1's capacity for initiating learning tasks, and C2's stronger performance ability to facilitate iterative refinement.

Platform Comparison: Skill Domain Coverage

Figure 3 visualizes expert ratings of C1's coverage of the five skill domains.

To aid interpretation, each bubble represents one expert's total score for a given skill domain—the larger the bubble, the higher the perceived adequacy of that skill within the AI-generated lesson plan (maximum possible score: 60).

Alt. Text: Bubble chart of expert skill ratings for C1 by skill type; bubble area ∝ score; engineering skills lowest on average.

The chart reveals a clear pattern: consistently larger bubbles in Conceptual ($\bar{x}$ = 49.45), Domain-Specific ($\bar{x}$ = 47.91), and Social-Emotional Learning Skills ($\bar{x}$ = 47.64) indicate strong and agreed-upon performance in these areas.

In contrast, the notably smaller bubbles in Engineering Skills ($\bar{x}$ = 37.45) reflect a significant weakness in addressing process-based, iterative design competencies.

This visual pattern corroborates expert comments attributing C1's engineering-skills deficit to its inability to simulate iterative, process-based activities.

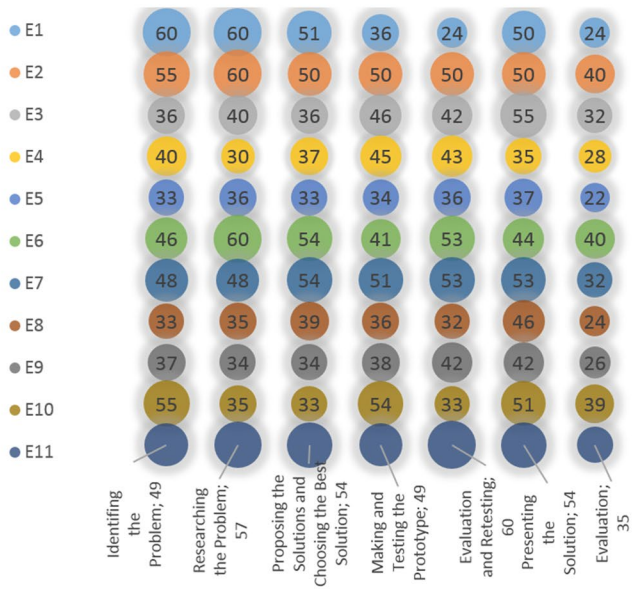


Fig. 1 Experts' ratings of the stages of the learning approach for C1. Note: $\bar{x}_{IP}$=44.73,$\bar{x}_{RP}$=45.00, $\bar{x}_{PSCBS}$=43.18, $\bar{x}_{MTP}$=43.64,$\bar{x}_{ER}$=42.55, $\bar{x}_{PS}$=47.00$\bar{x}$PS>$\bar{x}_{RP}$≈ $\bar{x}_{IP}$> $\bar{x}_{MTP}$≈ $\bar{x}_{PSCBS}$>$\bar{x}_{ER}$

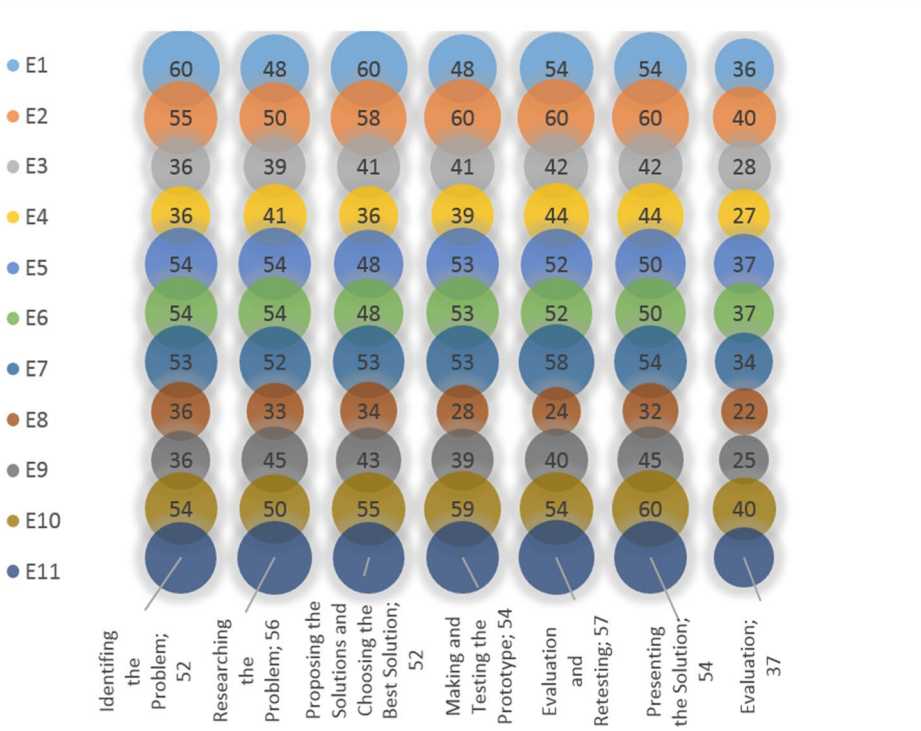


Fig. 2 Experts' ratings of the stages of the learning approach for C2. Note: $\bar{x}_{IP}$=46.00,$\bar{x}_{RP}$=46.55, $\bar{x}_{PSCBS}$=47.36, $\bar{x}_{MTP}$=47.45,$\bar{x}_{ER}$=48.45, $\bar{x}_{PS}$=49.09$\bar{x}$PS> $\bar{x}_{ER}$>$\bar{x}_{MTP}$ ≈ $\bar{x}_{PSCBS}$ > $\bar{x}_{RP}$≈ $\bar{x}_{PIP}$

Despite this limitation, C1's overall strong performance in core curriculum skills explains its favorability among some experts.

Figure 4 depicts expert ratings of C2's coverage across the five skill domains.

To aid interpretation, each bubble represents one expert's score for a given skill domain—the larger the bubble, the higher the perceived adequacy (maximum score = 60).

Alt. Text: Bubble chart visualising the score given by each expert to C2 for each of the skills investigated.

The large, consistent bubbles for Conceptual ($\bar{x}$ = 50.00), Domain-Specific ($\bar{x}$ = 50.72), and Literacy Skills ($\bar{x}$ = 49.45) indicate C2's comprehensive and reliable coverage of core curriculum competencies. While Engineering Skills ($\bar{x}$ = 41.45) still show smaller bubbles relative to other domains—highlighting a persistent challenge for both platforms—C2's performance represents a marked improvement over C1 ($\bar{x}$ = 37.45). This visual evidence supports expert observations that C2's education-focused design enables more effective skill scaffolding, though gaps in engineering design competencies remain.

Figure 5 compares the average skill domain scores for C1 and C2. C1 demonstrated slight advantages in Conceptual ($\bar{x}$=49.45) and Social-Emotional skills ($\bar{x}$=47.64), while C2 C2 showed higher average scores in Domain-Specific, Literacy, and Engineering skills. Notably, both platforms received their lowest scores in Engineering skills, confirming a shared limitation in addressing design-based competencies, though C2 showed a 4-point higher average score over C1 in this domain.

Alt. Text: Comparative bar chart showing the average expert scores for C1 and C2 across five skill domains. C1 (darker bars) shows slightly higher scores in Conceptual and Social-Emotional skills. C2 (lighter bars) shows higher average scores than C1 in Domain-Specific, Literacy, and Engineering skills. Both platforms received their lowest scores in Engineering Skills, though C2 shows a clear improvement over C1 in this domain.

Overall Quality Assessment

Table 3 presents the comparative overall evaluation of both AI-generated lesson plans across eight quality criteria. From this expert panel, C2 received numerically higher average scores than C1 across all eight criteria. The largest numerical differences were observed in 'Duration' (+1.4 points), 'Customization to context' (+1.2 points), and 'Attention grabbing' (+1.0 points).

While both platforms received moderately high average scores, C1's lowest averages were for 'Duration' (6.9) and 'Increasing interest in STEM' (6.9). C2's highest averages were for 'Customization to context' (8.9) and 'Value acquisition' (8.9). Given the small number of experts (n = 11), these differences should be interpreted as descriptive patterns rather than evidence of statistical superiority.

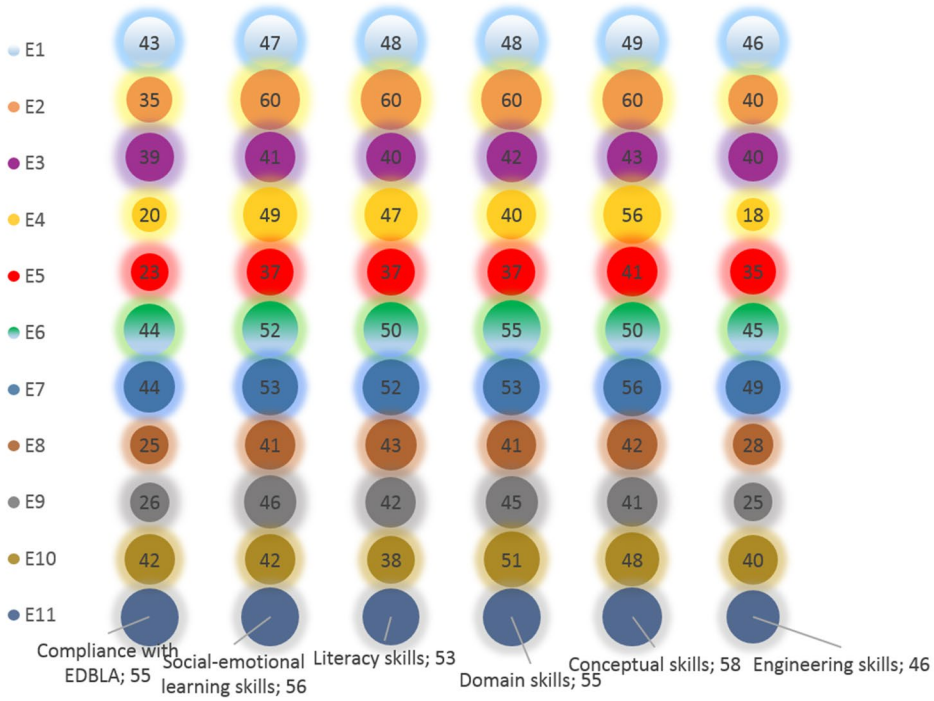


Fig. 3 Experts' assessment of C1 by skills. Note: $\bar{x}_{SELS}$=47.64,$\bar{x}_{LS}$=46.36, $\bar{x}_{DS}$=47.91, $\bar{x}_{CS}$=49.45,$\bar{x}_{ES}$=37.45$\bar{x}$CS>$\bar{x}_{DS}\approx \bar{x}_{SELS}> \bar{x}_{LS}>\bar{x}_{ES}$

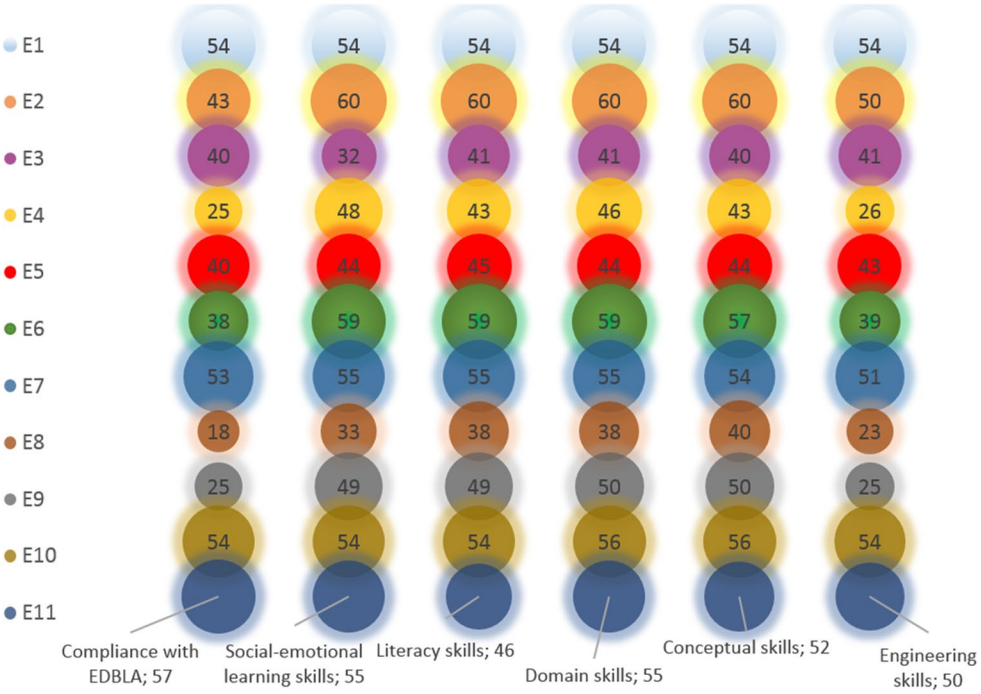


Fig. 4 Experts' assessment of C2 by skills. Note: $\bar{x}_{SELS}$=49.36,$\bar{x}_{LS}$=49.45, $\bar{x}_{DS}$=50.72, $\bar{x}_{CS}$=50.00,$\bar{x}_{ES}$=41.45$\bar{x}$DS> $\bar{x}_{CS}>\bar{x}_{LS}\approx \bar{x}_{SELS}> \bar{x}_{ES}$

Expert Preferences and Applicability

Analysis of individual expert preferences revealed clear patterns: Experts E2, E5, and E10 showed strong preference for C2 across most stages, while E4 demonstrated mixed preferences. Interestingly, E11 evaluated both plans similarly but ultimately preferred C1, specifically noting:

> "The first plan is more detailed and puts the student first. It encourages students to learn more. There is more emphasis on social-emotional skills… That's why I would choose the first plan." (E11).

This qualitative insight helps explain why some experts favored C1 despite C2's higher quantitative scores in

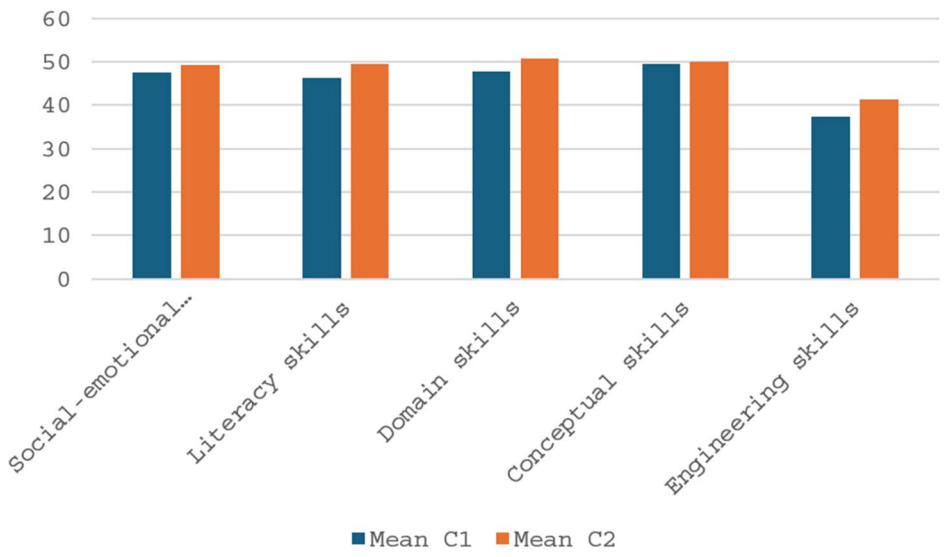


Fig. 5 Detailed comparison of the skills based on the points given

iterative stages. (See Supplementary Figure S1 for detailed expert-by-expert comparisons).

Table 4 shows total expert scores and final plan preference.

Notably, the preference split in Table 4 (7 experts preferring C2, 4 preferring C1, and 1 expressing no preference) indicates that although C2 received higher average scores in Table 2, expert opinions were not unanimous. This variability further cautions against overinterpreting small mean differences from a limited panel. This preference for C1, despite C2's higher overall scores, was primarily attributed to its stronger emphasis on social-emotional elements and student-centered design.

Furthermore, expert responses to the open-ended applicability question were analyzed thematically, resulting in three distinct categories (Table 5). The analysis revealed that while only 3 experts considered the plans directly 'Applicable', the majority (7 experts) endorsed conditional application with modifications ('Can be applied by correction'). Only one expert rejected both plans entirely ('Should not be applied'), primarily citing lack of originality:

> "These are copy-paste versions of existing MoNE plans… AI-generated plans lack novelty and fail to reflect authentic engineering design thinking." (E9).

Others emphasised adaptability:

> "The AI needs to be trained more on what engineering design-based learning is." (E4).

> "Such plans can be useful as drafts but not ready-made solutions." (E1).

In sum, while AI-generated plans hold promise in supporting skills-based instruction, they require teacher moderation for contextual and pedagogical depth.

Table 3 Comparative Overall Evaluation of AI-Generated Lesson Plans

Criterion	Mean C1	Mean C2	Difference (C2 - C1)
Learning outcome	7.6	7.9	+0.3
Grade level	8.1	8.7	+0.6
Duration	6.9	8.3	+1.4
Availability of materials	7.9	8.7	+0.8
Customization to context	7.7	8.9	+1.2
Attention grabbing	7.8	8.8	+1.0
Increasing interest in STEM	6.9	7.8	+0.9
Value acquisition	8.3	8.9	+0.6

Table 4 Analysis of individual expert preferences

Expert No	C1	C2	Preferred Plan
E1	287	333	C2
E2	245	347	C2
E3	303	311	C2
E4	258	267	C2
E5	231	292	C2
E6	338	348	C1
E7	339	357	C1
E8	245	209	C1
E9	253	273	None
E10	300	372	C2
E11	358	362	C1

Table 5 Thematic analysis of applicability assessments

Application Status	f
Applicable	3
Can be applied by correction	7
Should not be applied	1

Discussion

This study aimed to evaluate the pedagogical appropriateness of lesson plans generated by artificial intelligence (AI) platforms—one general-purpose and the other education-focused—within the framework of engineering design-based learning (EDBL) and the skill domains emphasized in the updated Turkish secondary science curriculum. In doing so, it explored the extent to which these platforms can support teachers in designing meaningful learning experiences, particularly through alignment with constructivist approaches and competency-based education goals.

One might ask why the study focused solely on the EDBL approach rather than using a more general instructional model. A possible concern could be that incorporating skills into the evaluation might obscure the effectiveness of the learning approach itself. However, it is important to acknowledge that lesson planning is not limited to the structuring of content or teaching techniques; it is inherently multidimensional and skill-oriented (Ndihokubwayo et al., 2022). As the new curriculum emphasizes the integration of conceptual, domain-specific, literacy, and social-emotional learning skills, lesson plans must reflect not only pedagogical structure but also support for holistic student development. Hence, this study adopted an integrated lens that simultaneously considers process (i.e., EDBL stages) and outcome (i.e., skills).

The findings show that while both AI platforms demonstrated an overall ability to produce instructional materials aligned with skill development, they were inconsistent in addressing the stages of the EDBL model. The general-purpose platform (C1) received relatively higher scores in the early stages of problem identification and research, whereas the education-focused platform (C2) performed better in the evaluation, prototyping, and solution refinement phases. This divergence likely stems from platform-specific affordances—C1 being optimized for open-ended content generation and C2 for structured pedagogical outputs (Davis & Lee, 2023).

Moreover, the platforms exhibited a strong tendency to generate well-structured final outputs ("Presenting the solution") but struggled with the iterative and cognitive depth required in EDBL. This is consistent with previous literature indicating that generative AI models often prioritize outcome-based formatting over process-based thinking (van den Berg & du Plessis, 2023). Notably, both platforms scored relatively lower in engineering skills compared to other competencies, indicating their limited capacity to replicate authentic design-based learning tasks such as failure analysis or prototype revision—elements central to the EDBL framework (English & King, 2015).

A particularly interesting pattern emerged in expert preferences: while C2 often received higher quantitative scores, some experts expressed a preference for C1. This suggests that expert judgment incorporates more than structural alignment—it also includes affective and contextual dimensions. For instance, one expert favored C1 for its stronger emphasis on student agency and social-emotional learning. This reinforces the argument that AI-generated plans must be evaluated not only for their technical fidelity but also for their potential to foster human-centered, inclusive learning environments (Carr, 2023; Rasul et al., 2023).

The issue of contextualization also emerged as a significant limitation. Experts frequently noted that neither platform adequately adapted lesson plans to the local sociocultural or classroom-specific context, even when prompted. While previous studies have highlighted AI's potential for content individualization (Gupta et al., 2023; Moreno-Guerrero et al., 2020), other scholars have emphasized persistent barriers to effective customization (Truong et al., 2023). The present study supports the latter position, suggesting that while the platforms offer surface-level alignment with curricula, they struggle to embed contextually grounded learning experiences—especially when user inputs are non-English or culturally specific.

Another limitation relates to linguistic access. The researchers' attempts to generate lesson plans using the platforms in their native language (Turkish) produced limited results, while switching to English yielded more pedagogically structured outputs. This finding reveals a digital equity issue: AI tools may disproportionately serve users with access to English-language proficiency, creating disparities in pedagogical quality across linguistic contexts.

Generalizability and Contextual Considerations

While findings are context-bound to the Turkish science curriculum, the observed tendencies in AI performance—particularly the trade-off between structure and iteration—are likely generalizable to similar competency-based contexts. While this study was conducted within the specific context of Turkey's 6th-grade biodiversity unit, the findings offer insights that may extend to other educational settings. The consistent patterns observed in AI platforms' strengths and limitations—particularly their proficiency in structured output generation versus challenges with iterative processes—likely transcend specific curricular contexts. The education-focused platform's stronger performance in supporting EDBL's iterative stages suggests that pedagogical training in AI systems can yield meaningful improvements, regardless of subject matter or grade level.

However, the cultural and linguistic specificity of our study highlights important boundary conditions. The platforms' struggle with Turkish-language prompts and local contextualization suggests that their performance may vary

significantly across different linguistic and cultural environments. The finding that both platforms achieved their lowest scores in engineering skills across all evaluations indicates a fundamental challenge in AI's ability to simulate complex, process-oriented learning tasks—a limitation that likely persists across diverse educational contexts.

Future research should explore whether similar patterns emerge in different curricular frameworks, grade levels, and cultural contexts. Comparative studies across countries with varying educational traditions and AI adoption levels would be particularly valuable for understanding how cultural and institutional factors mediate AI's educational utility.

Limitations and Future Research Directions

Several limitations of this study should be acknowledged. First, the reliance solely on expert evaluations, while providing valuable insights into pedagogical quality, means we cannot assess how these AI-generated plans would actually perform in classroom implementation. The absence of data on student learning outcomes or teacher experiences represents a significant gap that future research should address. Expert judgment, while essential for establishing pedagogical validity, cannot substitute for evidence of actual learning effectiveness.

Second, the study's focus on a single subject area (science) and grade level (6th grade) limits the generalizability of findings across disciplines and age groups. Different subjects may impose distinct demands on AI systems, and grade-level appropriateness involves developmental considerations that this study could not capture comprehensively.

Third, the evaluation timeframe was limited to immediate expert assessment without considering longitudinal aspects of lesson plan implementation. How teachers might adapt and refine AI-generated plans over time, and how students might respond to sustained exposure to AI-designed instruction, remain open questions.

Fourth, the necessity of translating AI outputs from English to Turkish introduces a potential source of variation. While the translation was performed carefully by domain experts, we cannot completely rule out that subtle pedagogical nuances may have been affected. Future research should prioritize AI platforms that support multiple languages natively to avoid translation-related artifacts. Additionally, the fact that Turkish-language prompts yielded poorer EDBL alignment suggests that current AI systems may have uneven performance across languages, a finding that has equity implications for non-English educational contexts.

Fifth, although blinding was applied, we cannot completely exclude the possibility that experts with extensive ChatGPT experience might have recognized certain stylistic features. Nevertheless, the education-focused platform (C2) received higher average scores on most criteria, suggesting that any residual familiarity bias did not systematically favor the general-purpose platform (C1).

Future research should triangulate these expert findings with classroom observations, student achievement data, and teacher implementation experiences to provide a more comprehensive understanding of AI-generated lesson plans' practical utility. Longitudinal studies examining how teachers integrate AI tools into their planning processes over time would be particularly valuable. Additionally, research exploring AI's capacity to support differentiated instruction and accommodate diverse learner needs would address important gaps in current understanding. Building on the preceding discussion, the following section concludes by summarizing the study's key theoretical and practical implications.

Conclusion

This study demonstrates that AI platforms can serve as valuable assistants in lesson planning but require significant teacher mediation and pedagogical expertise. The comparative analysis reveals distinct strengths and limitations between general-purpose and education-focused AI platforms, with the latter showing stronger alignment with sophisticated pedagogical frameworks like EDBL. The education-focused platform (C2) demonstrated stronger performance in supporting iterative learning processes and skill development, while the general-purpose platform (C1) excelled in generating socially engaging content and addressing affective learning dimensions.

Theoretical and Practical Contributions

Theoretical Contribution: This research contributes to the growing literature on AI in education by providing empirical evidence of how different types of AI platforms perform against established pedagogical models like EDBL and competency-based frameworks. The findings challenge simplistic narratives about AI's educational potential by revealing nuanced patterns of strengths and limitations across different aspects of instructional design.

Practical Implications: For educators, these findings emphasize that AI-generated lesson plans should be treated as starting points rather than finished products. The critical evaluation and adaptation of AI outputs emerge as essential teacher competencies in the AI era. Educational institutions should develop professional development programs that equip teachers with the skills to critically evaluate, adapt, and enhance AI-generated educational materials.

For AI developers, the results highlight the importance of building pedagogical knowledge into educational AI

systems. The stronger performance of the education-focused platform suggests that domain-specific training can significantly enhance AI's educational utility. Developers should prioritize deeper integration of learning science principles and pedagogical models into AI architectures.

Policy Recommendations and Future Directions

In light of these conclusions, several implications emerge for educational policy and practice:

- Educational institutions and teacher training programs should guide educators in the critical use of AI platforms, focusing not only on technical skills but also on pedagogical judgment and adaptation strategies.
- AI platform developers should integrate established pedagogical models such as EDBL into system design, enabling more coherent and process-oriented instructional outputs.
- Education ministries and regulatory bodies should provide guidance on the ethical, linguistic, and pedagogical use of AI, ensuring accessibility and preserving teacher autonomy while leveraging AI's potential benefits.
- Improving AI tools' adaptability to non-English contexts and diverse cultural frameworks is essential for equitable implementation across global educational systems.

Future research should: (1) examine AI-generated lesson plans in diverse cultural and linguistic contexts; (2) investigate classroom implementation and student learning outcomes through rigorous experimental designs; (3) explore how teacher professional development can enhance critical engagement with AI tools; and (4) develop more sophisticated evaluation frameworks that capture the complex interplay between AI capabilities and pedagogical requirements.

Ultimately, as AI continues to evolve, the most effective educational applications will likely emerge from thoughtful partnerships between AI capabilities and human pedagogical expertise. This study represents an initial step toward understanding how such partnerships might be structured to enhance rather than replace teachers' professional judgment and to support rather than undermine meaningful learning experiences.

Author Contributions All authors contributed the study conception, design, data analysis and writing of the manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors did not receive any support from funding agencies.

Data Availability Not applicable.

Declarations

Ethical Approval Ethics committee approval was obtained from Mersin University for the research.

Consent to Participate All participants voluntarily agreed to take part in the study.

Consent to Publish Not applicable.

Informed Consent All experts who participated in the study provided informed consent prior to their involvement.

Research Involving Human Participants and/or Animals Not applicable.

Competing interests No potential conflict of interest was reported by the author(s).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Carr, B. (2023). Revolutionizing education: Unleashing the power of Chat GPT/AI to empower educators. *Technology and the Curriculum: Summer 2023.* https://techandcurriculum.pressbooks.com/

Davis, R. O., & Lee, Y. J. (2023). Prompt: ChatGPT, create my course, please! *Education Sciences*, *14*(1), 24. https://doi.org/10.3390/educsci14010024

English, L. D., & King, D. T. (2015). STEM learning through engineering design: Fourth-grade students' investigations in aerospace. *International Journal of STEM Education, 2*(1), Article 14. https://doi.org/10.1186/s40594-015-0027-7

Gupta, P. K., Raturi, S., & Venkateswarlu, P. (2023). ChatGPT for designing course outlines: A boon or bane to modern technology. *SSRN*. https://doi.org/10.2139/ssrn.4386113

Hashem, R., Ali, N., El Zein, F., Fidalgo, P., & Khurma, O. A. (2024). Online collaborative writing in an online EFL writing class. *Research and Practice in Technology Enhanced Learning, 19*, Article 022. https://doi.org/10.58459/rptel.2024.19022

Karaman, M. R. (2024). Are lesson plans created by ChatGPT more effective? An experimental study. *International Journal of Technology in Education*, *7*(1), 107–127. https://doi.org/10.5281/zenodo.10646816

Lee, G. G., Mun, S., Shin, M. K., & Zhai, X. (2024). Confronting imminent challenges in humane epistemic agency in science

education: An interview with ChatGPT. *Science & Education*. https://doi.org/10.1007/s11191-024-00515-1

Moreno-Guerrero, A. J., López-Belmonte, J., Marín-Marín, J. A., & Soler-Costa, R. (2020). Scientific development of educational artificial intelligence in Web of Science. *Future Internet, 12*(8), Article 124. https://doi.org/10.3390/fi12080124

Ndihokubwayo, K., Byukusenge, C., Byusa, E., Habiyaremye, H. T., Mbonyiryivuze, A., & Mukagihana, J. (2022). Lesson plan analysis protocol (LPAP): A useful tool for researchers and educational evaluators. *Heliyon*, *8*(1), e08730. https://doi.org/10.1016/j.heliyon.2022.e08730

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). SAGE Publications.

Rasul, T., Nair, S., Kalendra, D., Robin, M., Santini, F. O., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, *6*(1), 41–56. https://doi.org/10.37074/jalt.2023.6.1.29

Truong, H., Nguyen, P., Cao, L., Nguyen, T., & Nguyen, P. (2023). Role of ChatGPT in Vietnamese education. *OSF Preprints*. https://doi.org/10.35542/osf.io/52smv

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

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.