Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Developing the AI-REFINE Process to Foster Pre-service Science Teachers' Competency: A Delphi Validation and Pilot Study

Do Huong Tra
Hanoi National University of Education
Nguyen Quang Linh
linhnq@tnue.edu.vn

Thai Nguyen University

Research Article

Keywords: Science teacher education, Pre-service science teachers, Generative AI, AI-REFINE, Co-design, Pedagogical Content Knowledge (PCK)

Posted Date: May 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9477558/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Generative AI offers transformative potential for science lesson planning but poses risks of passive use and scientific hallucinations for pre-service science teachers (PSTs), highlighting an urgent need for structured co-design frameworks in science teacher education. This study aimed to develop and content-validate the AI-REFINE process to foster PSTs' competency in co-designing science-integrated Science, Technology, Engineering, and Mathematics (STEM) lesson plans with AI. The research first utilized a two-round Delphi technique with nine science education experts to establish content validity for the six-step model (Request, Evaluate, Formulate, Iterate, Normalize, Export). Subsequently, to explore the model's practical feasibility, a six-week pilot study was conducted with 28 natural science PSTs using a one-group pre-test/post-test design. Lesson plans were assessed by two independent raters to ensure evaluation objectivity. The Delphi phase confirmed excellent content validity with an Scale Content Validity Index (S-CVI/Ave) of 0.97. The pilot study results indicated a statistically significant improvement in co-design competency scores (increasing from mean 5.77 to 7.80, $p < 0.001$). These initial findings suggest that AI-REFINE is a content-valid framework with promising potential for guiding PSTs to become active human-in-the-loop co-designers, thereby supporting pedagogical control and scientific integrity. Further experimental research is recommended to establish broader causal efficacy.

1. Introduction

1.1. The Context of the Generative AI Explosion in Education

In recent years, the landscape of science education has been fundamentally reshaped by the rapid emergence of Generative AI (GenAI). The launch of Large Language Models (LLMs) like ChatGPT and Gemini has ignited a critical debate specifically within the science teacher education community, focusing on the balance between technological innovation and scientific rigor (Baxmann et al., 2025; Fenta, 2025). This rise is viewed as a paradigm shift that necessitates a re-evaluation of how science teachers are trained to design complex learning environments (Đỗ, 2025; Wang et al., 2025). The rapid pace of GenAI development poses a significant challenge for PSTs, who often face a gap between AI's computational capabilities and their own pedagogical readiness to implement it effectively in inquiry-based settings. Educational organizations have recognized these challenges and issued urgent recommendations for the responsible use of AI in science and STEM contexts (Nguyen, 2025; Pandey et al., 2025). UNESCO’s “Guidance for Gen AI in education and research” (Holmes & Miao, 2023), further emphasizes a human-centered approach, which is vital for maintaining the teacher’s pedagogical control in science classrooms. Consequently, the central question for science teacher education has shifted: how to redefine the teacher's role as a facilitator of scientific inquiry in an era where machines can simulate reasoning and creation.

1.2. The Duality of AI in Lesson Design: Potentials and Challenges

In the context of science teacher education, GenAI clearly demonstrates its duality. On one hand, these tools offer immense potential to become effective pedagogical assistants for science educators. Studies have shown that GenAI can help science teachers significantly reduce administrative workload and lesson preparation time (Misawa et al., 2025; Wu & Carroll, 2025). In the role of a collaborator, GenAI can effectively support science lesson planning (Abbonizio et al., 2025; Alrwaished, 2024), provide creative ideation for inquiry-based activities, diversify scientific teaching materials, and build rich teaching scenarios grounded in real-world scientific phenomena (Wardak et al., 2024). By automating repetitive tasks, AI allows PSTs to reallocate their cognitive resources, focusing on higher-value aspects such as personal interaction with students and designing in-depth formative assessment strategies to measure specific scientific competencies.

On the other hand, these potentials come with serious pedagogical risks, especially when the users are PSTs, future teachers who lack practical experience and a solid theoretical foundation in specialized Science Pedagogical Content Knowledge (Science PCK). These risks include:

Scientific Hallucinations: GenAI tends to fabricate information, making scientifically incorrect claims and creating non-existent references (Misawa et al., 2025; Qian, 2025). This is a particularly dangerous risk in science education, a field requiring absolute accuracy of foundational knowledge and empirical evidence.

Algorithmic and Scientific Biases: AI models, trained on massive data from the Internet, may inadvertently reproduce and amplify unacknowledged biases regarding gender, race, and culture within the scientific community (Liu et al., 2024). Research on the psychosocial impacts of these biases has shown negative effects on diverse student groups (Lindell & Utterberg Modén, 2025).

Risk of passive use and decline in science pedagogical competency: This is the core risk. Lacking a structured guidance process, science PSTs are at high risk of falling into a state of passive use, performing copy-paste behaviors without critical thinking required for scientific inquiry (García-Carmona, 2025lıçkaya & Kic-Drgas, 2025). This situation not only deprives them of the opportunity to develop basic curriculum design skills but can also lead to a degenerative effect on professional judgment and their Science Pedagogical Content Knowledge (Science PCK) (Alrwaished, 2024) and the pedagogical competency of teachers (Ganesan & Morales, 2024).

The paradox is that the very tool designed to support science teachers actually risks hindering the development of their core competencies. Therefore, the problem lies not in the technology itself, but in the competency of the user to maintain pedagogical control over scientific content.

1.3. The Urgent Need for Co-design Competency for PSTs

To resolve the aforementioned paradox, teacher training institutions cannot choose the binary choice of banning or allowing the use of AI. Instead, an urgent need is set to equip PSTs with a new set of competencies so they can master this technology (Herro et al., 2025). There needs to be a fundamental shift in training objectives: from viewing PSTs as passive users of AI products to active co-designers

who collaborate with AI to create high-quality pedagogical products. This transition process requires PSTs to achieve a sequence of connected competencies (Mahat et al., 2023; Mejeh & Sarbach, 2025). The foundation is Critical AI Literacy, including knowledge, skills, and attitudes to critically evaluate AI outputs, identify biases, and ethical implications. Based on that, PSTs must apply the Human-in-the-Loop (HITL) principle (Penuel et al., 2022), an action philosophy emphasized by the U.S. Department of Education and many researchers, asserting that the educator must always be the central decision-maker. The goal of applying HITL is for PSTs to maintain pedagogical control, proactively setting the learning context and input parameters for AI, rather than being led by AI (Smith et al., 2025; Zeivots et al., 2025). Finally, the highest expression of this competency is the ability to co-design (Penuel et al., 2022), where PSTs and AI operate as partners, combining the computational power of machines with the refined pedagogical intelligence of humans.

1.4. The Research Gap

Although the need to foster co-design competency is urgent, current studies and intervention programs remain scattered and fragmented. Most literature to date has focused primarily on exploring perceptions, attitudes, or adoption barriers of PSTs regarding GenAI (Abate et al., 2022; Al-Nuaimi & Al-Ghamdi, 2022), without providing structured intervention models. Furthermore, many training programs are often narrowed down to isolated technical skills such as Prompt Engineering (Ba et al., 2023; Nguyễn et al., 2023). From here, a dual gap clearly emerges: First, the theoretical gap: There is a lack of comprehensive, structured, and especially content-validated models or processes to foster this in-depth co-design competency. Second, the practical gap: Even when theoretical models are proposed, there is a shortage of empirical evidence regarding their application and initial effectiveness in science teacher education contexts. This gap becomes even more apparent in specific fields like STEM education, which requires a delicate balance between scientific accuracy and pedagogical creativity.

1.5. Statement of Purpose and Research Questions

To address the identified dual gap, this paper presents the results of a mixed-methods study designed in two sequential phases: Phase 1: Development and Validation. We (1) develop a structured 6-step process named AI-REFINE (standing for Articulate Idea, Request, Evaluate, Formulate, Iterate, and Normalize & Export). Then, we (2) validate the completeness, rationality, and feasibility of this process using the Delphi Method. A panel of nine experts participated in multiple rounds to refine and validate the process. Phase 2: Feasibility and Initial Effectiveness Evaluation (Pilot Study). After obtaining the validated process from Phase 1, we (3) explore the practical feasibility and potential effectiveness of the AI-REFINE process through a one-group pilot study. We examine whether introducing PSTs to the AI-REFINE process shows positive trends in enhancing their STEM lesson plan co-design competency.

This paper will focus on answering the following research questions:

1. RQ1 (Phase 1, Delphi): What are the essential components (steps and connotations) of the process for fostering STEM lesson plan co-design competency with AI for PSTs (the AI-REFINE process)?
2. RQ2 (Phase 1, Delphi): What is the level of consensus among experts (quantified by CVI) regarding the necessity, clarity, and feasibility of the AI-REFINE process?
3. RQ3 (Phase 2, Pilot Study): To what extent does the AI-REFINE training pilot study show potential in improving the STEM lesson plan co-design competency of PSTs?
4. RQ4 (Phase 2, Pilot Study): What practical advantages and difficulties do PSTs encounter when applying the AI-REFINE process in an initial trial?

By introducing a process that is theoretically content-validated and accompanied by initial empirical evidence from a pilot study, this research provides a foundational intervention framework for teacher training institutions, contributing to filling the existing gap between technological potential and the practical competency needs of future teachers.

2. Literature Review

2.1. Generative AI in Science Teacher Education: Potentials and Pedagogical Risks

The integration of Generative Artificial Intelligence (GenAI) and LLMs into educational settings has sparked a paradigm shift, particularly in the landscape of science teacher education. Unlike traditional educational technologies that function primarily as static executor tools, GenAI acts as an interactive agent capable of processing complex scientific phenomena and generating dynamic instructional content. Recent studies strictly situated within science education highlight the substantial potential of GenAI to assist PSTs in developing inquiry-based lesson plans, modeling real-world scientific scenarios, and scaffolding complex STEM concepts for diverse learner profiles (Kılıçkaya & Kic-Drgas, 2025; Wang et al., 2025; Wu & Carroll, 2025). By offloading routine administrative tasks, AI allows science educators to allocate more cognitive resources toward facilitating hands-on scientific investigations and designing formative assessments.

However, the deployment of GenAI in science education introduces significant pedagogical risks, especially for novice teachers who are still developing their foundational Science Pedagogical Content Knowledge (Science PCK). The most critical risk is the phenomenon of “scientific hallucinations.” Operating on probabilistic prediction rather than factual awareness, LLMs frequently generate plausible but factually incorrect scientific claims, fabricate empirical data, or create non-existent citations. In science education - a discipline fundamentally grounded in empirical evidence and rigorous methodology - such hallucinations can substantially compromise instructional integrity. Furthermore, AI models often reproduce implicit societal biases regarding scientific discoveries, potentially marginalizing diverse student identities in STEM classrooms. Consequently, there is an overarching concern that without structured guidance, PSTs may adopt a passive “copy-paste” approach, unquestioningly accepting AI outputs and bypassing the critical cognitive processes required to design authentic

scientific inquiries. This passive reliance threatens to erode their emerging pedagogical competencies and professional judgment.

2.2. Pedagogical Content Knowledge and Teacher Agency in the AI Era

To navigate the dual nature of GenAI, science teacher education must emphasize the development of robust Pedagogical Content Knowledge (PCK) and teacher agency. Shulman's foundational concept of PCK - the intersection of deep subject matter knowledge and pedagogical strategies - remains the critical anchor for effective science instruction. In the contemporary landscape, this framework has evolved into Technological Pedagogical Content Knowledge (TPACK) and, more recently, AI-specific competency models tailored for educators (Shulman, 1987). For science PSTs, Science PCK acts as the essential “professional filter” required to evaluate AI-generated content. Without a deep conceptual understanding of scientific principles, PSTs cannot accurately identify scientific hallucinations or ensure that AI-suggested activities align with rigorous inquiry-based learning standards (Jang & Chen, 2010; Kurt & Çakıroğlu, 2025; Schiering et al., 2023).

Closely tied to Science PCK is the concept of teacher agency, defined as the capacity of educators to act purposefully and constructively direct their professional practice within complex ecological conditions. As GenAI assumes more active roles in instructional design, maintaining teacher agency becomes synonymous with exercising “pedagogical control”. HITL philosophy, which dictates that the human educator must remain the central decision-maker and critical evaluator in any AI-assisted educational process. For science teachers, exercising agency means proactively setting the scientific learning objectives, carefully engineering prompts to constrain AI outputs within scientifically valid parameters, and rigorously evaluating the generated lesson components. Fostering this agency requires a shift in teacher preparation programs: PSTs must be trained not merely as passive consumers of AI tools, but as critical curators who utilize technology to enhance human pedagogical reasoning.

2.3. Co-design Approaches in Teacher Training Programs

The ultimate goal of integrating AI into science teacher preparation is to elevate PSTs to the role of active co-designers. Co-design in the educational context traditionally refers to the collaborative development of curricula among multiple stakeholders. When applied to human-AI interaction, co-design represents a synergistic partnership where the computational efficiency and divergent ideation capabilities of AI are seamlessly merged with the contextual awareness, ethical judgment, and pedagogical empathy of the human teacher.

Recent literature exploring human-AI co-design suggests that structured collaborative processes yield higher quality instructional materials compared to purely human-generated or AI-generated outputs (Abbonizio et al., 2025; Phạm et al., 2025; Smith et al., 2025). However, a significant gap remains in the literature regarding specific applications within science and STEM education. Most existing empirical studies on AI in teacher training either focus on general education disciplines or stop at surveying PSTs'

attitudes and perceived barriers, lacking actionable intervention models. There is a distinct lack of comprehensive, structured pedagogical frameworks designed specifically to guide science PSTs through the iterative process of co-designing inquiry-based STEM lessons. While broad principles like iterative design thinking and backward design are widely acknowledged as effective in science curriculum development, their explicit integration into actionable, step-by-step AI co-design protocols remains underexplored. Therefore, there is an urgent need for research that proposes theoretically grounded frameworks for human-AI co-design in science education and provides empirical evidence of their initial effectiveness in authentic teacher preparation contexts.

3. Methodology

3.1. Theoretical Foundations and Research Setting

The theoretical foundation for this study is built upon the integration of HITL principles and the development of specialized pedagogical expertise within science teacher education. HITL is the central interaction framework in the AI era, asserting that humans must always be the core decision-makers in the system (Kılıçkaya & Kic-Drgas, 2025). In science teacher training, this means professionalizing the role of PSTs, requiring them to continuously verify, justify, and reflect on scientific results generated by AI (Baxmann et al., 2025; Liu et al., 2024). All steps of AI-REFINE, especially Evaluate and Iterate, are the operation of the HITL philosophy, ensuring AI is a supplementary tool and never replaces the human scientific judgment.

If HITL is the interaction philosophy, then Pedagogical Control is the goal of that interaction within the science classroom. Researchers warn of the risk of teachers losing pedagogical control when AI becomes too autonomous (Fenta, 2025; Misawa et al., 2025). This control is the educator's ability to establish goals, contexts, and scientific pedagogical values of the lesson. The ability to maintain this control depends on the teacher's pedagogical reasoning foundation, particularly their Science PCK. Shulman's theory of PCK defines the teacher's core knowledge as a special amalgam of content and pedagogy, helping them know how to express complex scientific content in a way that is most understandable to learners (Shulman, 1987). Inheriting this, the Technological Pedagogical Content Knowledge (TPACK) framework emphasizes the complex intersection of Technology, Pedagogy, and Content (Kurt & Çakıroğlu, 2025). In the GenAI context, PSTs must use their Science PCK and TPACK to exercise pedagogical control. Newer competency frameworks such as GenAI-TPACK or Artificial Intelligence Agent – Pedagogical Content Ethical Knowledge (AIA-PCEK) are also being developed to emphasize this role (Azcárate, 2024; Phạm et al., 2025). Step 1 (Articulate Idea) and Step 3 (Evaluate) of the AI-REFINE process are specifically designed for science PSTs to exercise pedagogical control: Step 1 requires them to establish scientific pedagogical parameters before using AI, and Step 3 requires them to act as a gatekeeper to validate output quality based on their Science PCK/TPACK.

Once control is mastered, science PSTs can shift from a passive control role to active co-design (Penuel et al., 2022; Smith et al., 2025). Co-design is a collaborative model where teachers and AI together create

educational products, leveraging the strengths of both (Wardak et al., 2024). This is the theoretical foundation for Step 4 (Formulate) of the AI-REFINE process, where science PSTs do not copy-paste but must integrate ideas from AI and add human pedagogical value to scientific concepts. This co-design ability is closely related to Teacher Agency. Accordingly, agency is an achievement emerging from the interaction between individual effort and surrounding ecological conditions. A training process like AI-REFINE is an ecological condition designed to activate and enable pre-service science PSTs to exercise their agency in the new technological context of science education.

The 6-step structure of AI-REFINE is built upon two widely recognized process frameworks applied to science curriculum development. Step 1 (Request & Articulate) and Step 6 (Export) are directly built on the philosophy of Backward Design to ensure the alignment of science learning outcomes (Wiggins & McTighe, 2005). The core principle of Understanding by Design (UbD) is to start the design process from desired results, i.e., science-specific standards and learning objectives, before planning activities. Gen AI is essentially weak at this; it can generate creative activities but they are often not aligned with specific science pedagogical goals. Therefore, Step 1 of AI-REFINE forces PSTs to perform the most important step of UbD: clearly defining scientific inquiry goals and output standards before seeking AI assistance. Similarly, Step 6 acts as the final quality control step of UbD, requiring PSTs to review the absolute alignment between Science Objectives - Assessment - Activities to maintain pedagogical rigor

Step 4 (Iterate) of AI-REFINE is built based on Iterative Design Thinking, which is crucial for developing innovative science instructional strategies. Design Thinking is a human-centered problem-solving process characterized by non-linearity and iteration (Gelerstein et al., 2016; Jacobson et al., 2015). Science lesson design is not a linear once-and-done process; it is a rapid prototyping cycle where ideas for lab experiments or scientific investigations are constantly refined. Gen AI is particularly useful in supporting this iterative cycle by generating diverse instructional models. Step 4 (Iterate) encourages science PSTs to use AI to quickly create multiple prototypes of a scientific activity, compare them, and fail fast in a safe environment, thereby continuously improving the scientific accuracy and engagement of the pedagogical product.

Moving from Phase 1 (theoretical validation) to Phase 2 (empirical verification) requires a shift in rationale: from model building to measuring the effectiveness of that model in the context of science teacher education. To affirm that the AI-REFINE process truly fosters competency for PSTs, reliable and valid measurement tools are needed. Measuring a complex competency structure like science lesson co-design competency is inherently challenging, as it is not just general pedagogical knowledge but also specialized practical skills required to create high-quality science instruction. Therefore, the rationale for measurement needs to be based on two main pillars: product assessment of science lesson plans and science-specific knowledge assessment.

In science teacher training, the most common and effective method to assess instructional design competency is through analyzing output products, i.e., Science Lesson Plans compiled by students (Alrwaished, 2024; Chen & Zhang, 2019). To make this assessment objective and scientific, a science-

specific Rubric is used. In the context of this study, the Rubric tool set used to assess PSTs' lesson plans in Phase 2 is designed to accurately measure the goals that the AI-REFINE process aims for within science education (Fensham & Rennie, 2013; Hairida, 2016). The criteria in the Rubric reflect the core steps of the process:

+ Criterion on Science Pedagogical Soundness: Measures the PST's ability to clearly define scientific inquiry goals, context, and analyze output standards. This criterion corresponds directly to Step 1 and Step 5, especially the requirement for Alignment between Science Objectives - Activities - Assessment.

+ Criterion on Scientific Accuracy & Ethics: Measures the PST's critical competency when validating results from AI, including the ability to detect false scientific claims, logical errors, and unacknowledged biases in scientific contexts. This is the direct result of Step 2 (Evaluate).

+ Criterion on Human Value in Science Teaching: Measures the PST's ability to add human value to the product, demonstrating scientific wisdom and promoting student agency through inquiry-based learning — factors that AI often performs poorly. This criterion measures the success of Step 3 (Formulate).

Besides the final product, an effective intervention program also needs to enhance the specialized knowledge foundation for learners (Al-Nuaimi & Al-Ghamdi, 2022; Nguyễn Thị Diễm & Lê Danh, 2021). The pre-test/post-test design is a classic experimental method to quantify changes in science pedagogical knowledge. In this study, the test measures PSTs' foundational knowledge of core concepts such as AI risks like scientific hallucinations, principles of Backward Design for science curriculum, or prompting techniques for scientific ideation. Thus, Phase 2 utilizes the rationale of educational measurement, using a mixed toolset (Product Assessment Rubric and Science Knowledge Test) to comprehensively evaluate the intervention effectiveness of the AI-REFINE process in science teacher education.

3.2. Phase 1: Process Development and Validation (Delphi Study)

This study was designed using a mixed-methods sequential explanatory design, implemented through two main phases. Phase 1, aimed at development and validation, utilized the Delphi technique to construct and validate the AI-REFINE process specifically for science-integrated STEM lesson planning. The rationale for selecting the Delphi method is grounded in the nascent nature of GenAI application in science education, where established pedagogical frameworks and standardized evaluation criteria are currently lacking. In such pioneering contexts, achieving consensus among domain experts is a methodologically robust approach to establish the foundational content validity of a newly designed instructional process. It is important to acknowledge that while the Delphi study rigorously establishes content validity - ensuring that the proposed steps are logically sound, necessary, and theoretically aligned with science pedagogy - it serves as the requisite first stage of validation. Establishing other psychometric properties, such as construct or predictive validity, necessitates longitudinal, large-scale empirical implementations, which falls outside the scope of this initial model development and pilot

phase. To implement Phase 1, a panel of 09 experts was purposively selected. Selection criteria included: (1) holding a PhD or Associate Professor title; (2) having in-depth expertise in Science Education, Didactics, or Science-focused STEM Education; and (3) having 10 years or more of professional experience. These 09 experts came from three key science teacher training institutions in Vietnam. The Delphi process was conducted from September 6, 2025, to September 28, 2025, through 02 rounds. In Round 1 (Qualitative), the research team sent an open-ended questionnaire to collect in-depth opinions to critique and construct the initial draft of the process. In Round 2 (Quantitative), experts evaluated the necessity, clarity, and feasibility of the refined process using a 4-point Likert scale. The results confirmed an high level of consensus with an S-CVI/Ave reached 0.97 (Denzin & Lincoln, 2018), ensuring the content validity of the process within the science teacher education context.

3.3. Phase 2: Feasibility and Initial Effectiveness Evaluation (Pilot Study)

Phase 2 was designed as a pilot study to evaluate the practical feasibility and initial effectiveness of the AI-REFINE process in an authentic teacher education setting. Conducted from October 6, 2025, to November 13, 2025, this phase employed a one-group pre-test/post-test design. The participants were a purposive sample of 28 third-year PSTs majoring in Natural Science Education at Thai Nguyen University of Education. Given the nascent stage of the AI-REFINE model following its Delphi validation, a pilot study with a single cohort was deemed appropriate to gather initial empirical data and identify practical implementation challenges before scaling to a larger, controlled experimental design. The experimental subjects were 28 PSTs majoring in Bachelor of Natural Science in their third year at the University of Education - Thai Nguyen University. The pedagogical intervention lasted 6 weeks, integrated into the content of the Science Pedagogy module approved by the university. In Week 1, students were briefed and took the pre-intervention test. In the following 4 weeks, the instructor guided PSTs to apply the AI-REFINE process to practice designing science-focused STEM lesson plans. In Week 6, students took the post-intervention test and submitted their final products. To evaluate effectiveness, the research team constructed two independent tests based on science pedagogical knowledge specification matrix. These tests were content-validated by 03 experts and piloted on 52 PSTs to confirm reliability and equivalence (Mustafy & Rahman, 2024; York, 2022). All student works were graded independently by 02 raters who were lecturers experienced in science and STEM education. To ensure the reliability and objectivity of the product assessment, an inter-rater reliability analysis was conducted. The agreement between the two independent raters was measured, yielding an Intra-class Correlation Coefficient (ICC) of 0.83 (or Cohen's Kappa of 0.79), indicating a strong level of agreement in evaluating the PSTs' science lesson plans. Discrepancies in scoring were resolved through discussion until a consensus was reached. Official experimental data was analyzed using a Paired-Samples T-test on SPSS 26.0 to compare the difference in science co-design competency scores before and after the intervention, with the magnitude of impact calculated through Cohen's d index.

4. Research Results

4.1. Phase 1 Results: Development and Validation of the Process (Answering RQ1, RQ2)

Qualitative Results from the First Delphi Round

In Round 1, the research team received in-depth feedback from 09 science education experts. Content analysis results showed high consensus on the urgency and overall logical framework of the process within science teacher education. However, experts also pointed out significant limitations needing adjustment to ensure scientific accuracy and terminological consistency.

The main opinions were synthesized into three themes. Theme 1: Resolving confusion regarding terminology and naming. This was the issue receiving the most attention. Experts assessed that naming Step 1 as “A.I.” (Articulate Idea) in a process about Artificial Intelligence was inappropriate, easily causing misunderstanding that this was a step performed by AI rather than the science teacher. Experts emphasized this was a terminology error needing immediate change. Additionally, naming Step 6 as “N.E.” was evaluated as lacking formal consistency. Solution: The research team decided to keep the overall name “AI-REFINE” for brand recall but restructured the names of the internal steps to form a meaningful acronym, R-E-F-I-N-E, ensuring consistency and professionalism in science teacher training.

Theme 2: Structural refinement (Merging and Splitting steps). Regarding logic, experts proposed merging the Idea step and Request step, as in a co-design model, interaction with AI should occur from the beginning to clarify scientific ideas. Conversely, regarding the old Step 6, strong arguments suggested separating the “Normalize” action into an independent quality control checkpoint to ensure alignment with science pedagogical goals. Solution: Incorporating these opinions, the new process merged the idea presentation into the initial request step, while splitting the old final step into two separate steps: “Normalize” and “Export” to clarify the focus of each action within the science curriculum development process.

Theme 3: Supplementing science-specific connotations. Experts emphasized the need to add specific evaluation criteria for Step 2 (Evaluate), including scientific accuracy, technical feasibility in laboratory contexts, and ethical/social risks inherent in science education. Simultaneously, the non-linearity of the Iterate step was also clarified: although presented sequentially for teaching convenience, the operational nature allows science PSTs to return to previous steps at any time to refine their scientific instructional models.

Conclusion and AI-REFINE Model (Version 2.0): Based on the synthesis and absorption of the above opinions, the research team constructed the finalized version of the process for Round 2. The process kept the name AI-REFINE, but the steps were adjusted to 6 new steps corresponding to the letters R-E-F-I-N-E as shown in Table 1.

Table 1
Structure of the AI-REFINE Process (Finalized Version after Round 1)

Step	Step Name (New)	Connotation and Changes from the Draft
1	**R (Request & Articulate)**	**Merging old Steps 1 & 2.** PSTs define objectives and context (Articulate) right during the process of constructing the initial prompt (Request) to interact with AI. This thoroughly resolves the confusion caused by the old naming "A.I."
2	**E (Evaluate)**	**(Promoted from old Step 3).** Retains the core role as a pedagogical "filter." Adds evaluation criteria regarding scientific accuracy (STEM) and ethics based on expert feedback.
3	**F (Formulate)**	**(Promoted from old Step 4).** Focuses on "co-design" activities: integrating results from AI and adding educational value and human practical context.
4	**I (Iterate)**	**(Promoted from old Step 5).** Implements iterative design thinking. PSTs return to step R or E to refine the product. This position ensures process flexibility before finalizing the product.
5	**N (Normalize)**	**(Separated from old Step 6).** Focuses specifically on reviewing the "Alignment" between Objectives - Activities - Assessment and compliance with professional regulations (such as Official Letter 5512).
6	**E (Export)**	**(Separated from old Step 6).** The final technical step to package the product, finalize formatting, and cite AI sources transparently (following academic integrity recommendations).

The R-E-F-I-N-E model ensures tight science pedagogical logic, overcoming terminology errors from Round 1, and was ready for quantitative survey in Round 2.

Quantitative Results from the Second Delphi Round

After restructuring the process, the research team conducted the Round 2 survey to quantitatively determine the level of expert consensus on this new version. The survey included 15 detailed evaluation criteria on naming, logical structure, connotation of each step, and overall feasibility. Results from 09 experts showed very high consistency in the panel, with the S-CVI/Ave reaching 0.97, exceeding the recommended threshold (Table 2). This confirms that the R-E-F-I-N-E process achieved the necessary reliability and content validity for experimentation.

Table 2
Evaluation Results of the Content Validity Index (CVI) of the R-E-F-I-N-E Process (Round 2) (N = 9 experts)

Item Code	Evaluation Criteria (Summary)	No. of Experts in Agreement (Score 3 or 4)	I-CVI	Evaluation Result
A1	Appropriateness of R-E-F-I-N-E naming and terminology	9/9	1.00	Valid
A2	Logical sequence of the 6 steps	9/9	1.00	Valid
B1.1	Step 1 (R): Necessity of merging steps	9/9	1.00	Valid
B1.2	Step 1 (R): Clarity of connotation	9/9	1.00	Valid
B2.1	Step 2 (E - Evaluate): Necessity	9/9	1.00	Valid
B2.2	Step 2 (E): Clarity (STEM criteria)	9/9	1.00	Valid
B3.1	Step 3 (F - Formulate): Necessity	9/9	1.00	Valid
B3.2	Step 3 (F): Clarity	9/9	1.00	Valid
B4.1	Step 4 (I - Iterate): Necessity	9/9	1.00	Valid
B4.2	Step 4 (I): Clarity (Non-linearity)	8/9	0.89	Valid
B5.1	Step 5 (N - Normalize): Necessity (Step separation)	8/9	0.89	Valid
B5.2	Step 5 (N): Clarity (Alignment)	9/9	1.00	Valid
B6.1	Step 6 (E - Export): Necessity	9/9	1.00	Valid
B6.2	Step 6 (E): Clarity (Ethics/Citation)	9/9	1.00	Valid
C1	Feasibility of application for PSTs	8/9	0.89	Valid
Total	**Average (S-CVI/Ave)**		**0.97**	**Excellent**

Detailed analysis of Item-Level Content Validity Indices (I-CVI) showed omplete consensus from experts for strategic adjustments after Round 1. Specifically, criteria regarding naming and the new terminology system achieved an absolute I-CVI, proving that renaming steps to match a meaningful acronym resolved the terminology confusion issue. Similarly, the logical sequence arrangement from step R to step E also achieved 100% consensus.

Delving into the connotation of each step, adjustments regarding merging and splitting steps received positive evaluations. The new Step 1 “R”, merged from goal definition and prompting techniques, achieved an strong support for both necessity and clarity. Notably, the decision to split the old quality control step into two independent steps, Step 5 “N” and Step 6 “E”, was also thoroughly validated. Results showed high consensus on the necessity of the Normalize step and strong support on the

connotation of Alignment in this step. Regarding applicability feasibility, results recorded strong consensus. Therefore, the Delphi process concluded at Round 2.

The AI-REFINE Model (Final Version)

Based on the strong support results from the expert panel in Round 2, the process for fostering STEM lesson plan co-design competency with AI was officially established with the name AI-REFINE. This is a meaningful acronym representing six continuous thinking and action steps: Request & Articulate, Evaluate, Formulate, Iterate, Normalize, and Export. This model not only provides a technical guide but also establishes a pedagogical thinking framework, ensuring students always maintain a central role (human-in-the-loop) in the interaction process with technology (Fig. 1).

The process begins with the “Pedagogical Input” phase with two steps R and E. Unlike common approaches, step R does not stop at prompting techniques but requires students to clearly Articulate pedagogical goals and context right during the interaction setup. Step E activates critical thinking, requiring students to act as a “professional filter”. The next phase is “Co-design and Optimization”, including steps F and I. Step F is the center of creativity, where students implement a “selective integration” strategy, adding human value to create the draft lesson plan. Step I introduces iterative design thinking, encouraging continuous product improvement. The process ends with the “Quality Assurance and Finalization” phase with two separate steps N and E. Step N acts as a quality control checkpoint for Alignment check, and Step E involves finalization of format and academic ethics, specifically transparent citation of AI contributions.

4.2. Phase 2 Results: Pedagogical Experiment Effectiveness

Reliability Assessment and Descriptive Statistics

Before conducting comparative tests, the reliability of the scale and data was checked through Cronbach's Alpha coefficient. On the official experimental dataset, the reliability analysis result for the competency assessment scale was 0.929, confirming the internal consistency of the assessment tool.

Table 3
Summary of Comparison Results of Competency Assessment Scores Before and After Intervention

Test	N	Mean	Standard Deviation (SD)	t	df	Sig. (2-tailed)	Cohen's d
Pre-test	28	5.77	1.07	-18.48	27	< 0.001	3.49
Post-test	28	7.80	1.18				

Descriptive statistics results in Table 3 showed a clear improvement in students' competency scores after the fostering process. Specifically, before applying AI-REFINE, the average STEM lesson plan design competency score of students only reached a moderate level of 5.77. After 6 weeks of experimentation

with the process support, the average score increased to 7.80. The strong support increase was 2.03 points on a 10-point scale.

This change is visualized in Fig. 2, clearly indicating the increase in average scores, while also showing that data dispersion represented by error bars is equivalent at both time points, proving that progress occurred uniformly across the student group.

Hypothesis Testing on Impact Effectiveness

To confirm whether the difference in scores was statistically significant, a Paired-Samples T-test was performed. Results in Table 3 indicated a strong correlation between input and output competencies with a correlation coefficient $r = 0.87$ ($p < 0.001$). More importantly, the t-test result showed a highly statistically significant difference between pre-test and post-test scores with $t(27) = -18.48$, $p < 0.001$. The p-value is significantly smaller than the standard significance level of 0.05, indicating a statistically significant difference between the pre-test and post-test scores. While these initial findings suggest that applying the AI-REFINE process has the potential to support students' STEM lesson plan co-design competency, definitive causal inferences cannot be drawn due to the absence of a control group in this pilot design. Additionally, to evaluate the magnitude of the impact, the effect size was calculated. Given the high correlation between pre-test and post-test scores ($r = 0.870$, $p < 0.001$), calculating the standard Cohen's d_z using the standard deviation of the differences yields an inflated value. To provide a more conservative and interpretable estimate for educational research, Cohen's d_{av} was calculated using the average standard deviation of both time points. The result indicated a large effect size (Cohen's $d_{av} = 1.80$). This substantial effect highlights the promising potential of the AI-REFINE process; however, given the pilot nature of the study with a small sample size (N = 28) and the absence of a control group, this magnitude should be viewed as preliminary

Analysis of Practical Challenges

Besides positive quantitative results, the observation process and analysis of student design portfolios also revealed practical challenges when operating the process. First is the challenge of critical thinking at the Evaluate step. Although scores increased, many students reflected confusion when having to evaluate results from AI, especially detecting subtle errors in in-depth STEM knowledge. This shows that students' foundational professional competency remains a prerequisite factor. Second is the time pressure at the Iterate step. Due to the iterative nature of the process, students spent more time compared to traditional lesson planning or passive copying from AI. Having to return to the Request step to optimize prompts sometimes caused hesitation to revise in a segment of learners. However, these challenges themselves are evidence that students truly engaged in a deep co-design process rather than passive reception.

5. Discussion

5.1. Discussion on the Validity of the Process (RQ1 & RQ2)

Research results from Phase 1 confirmed the solid content validity of the AI-REFINE process within the specific context of science teacher education. With an S-CVI/Ave index reaching 0.97, this process far exceeds the recommended threshold to ensure the reliability of a new pedagogical intervention (Denzin & Lincoln, 2018). The high consensus from the expert panel not only validated the 6-step R-E-F-I-N-E structure but also reinforced the UNESCO view that applying AI in science education requires tightly structured guidance frameworks rather than leaving learners to navigate largely uncontrolled environments on their own.

The shift in focus from technical operations to pedagogical thinking steps such as Request & Articulate and Evaluate reflects consistency with modern educational theoretical frameworks. Experts particularly appreciated the integration of critical thinking into the very first step of the process. This aligns with the Human-in-the-Loop model emphasized (Alsaleh, 2020, p. 202; Lê & Nguyễn, 2025). By requiring students to set pedagogical parameters before interacting, the process restores science-specific pedagogical control to learners, addressing concerns about science teachers being led by AI or losing autonomy in lesson design (Kılıçkaya & Kic-Drgas, 2025; Misawa et al., 2025). Furthermore, separating the Normalize step to check alignment received omplete consensus in the Delphi study. This demonstrates the process's compatibility with the core principles of Understanding by Design, requiring absolute consistency between objectives, assessment, and teaching activities (Wiggins & McTighe, 2005).

5.2. Discussion on the Effectiveness of the Process (RQ3)

Pilot results from Phase 2 provided initial quantitative evidence regarding the feasibility and potential of the process in supporting the co-design skills of PSTs. The statistically significant difference in scores ($t(27) = -18.48$, $p < 0.001$) along with a large effect size (Cohen's $d_{av} = 1.80$) indicates a promising positive trend in fostering science lesson design competency, which aligns with recent discourse on the potential of AI support in science teacher education. This competency improvement can be explained through the operating mechanism of each step.

First, the fact that students achieved higher scores on criteria regarding scientific accuracy is due to the barrier of the Evaluate step. In a context where Gen AI frequently creates hallucinations and misleading information, requiring students to validate scientific information before use significantly eliminated errors in scientific concepts in lesson plans (Alrwaished, 2024; Phạm et al., 2025). Second, the increase in creativity scores is the result of the Formulate step. Instead of passive use like copy-paste, this process forces science students to act as active co-designers. This result aligns with findings asserting that when placed in an active position, science teachers can combine the computational power of AI with their specialized pedagogical intelligence to create products better than the individual capabilities of either side (Penuel et al., 2022; Wardak et al., 2024). Finally, iterative design thinking in the Iterate step helped students continuously optimize their inquiry-based products, transforming science lesson planning from a static task into a dynamic and adaptive constructive process.

5.3. Discussion on Practical Challenges (RQ4 and Expert Opinions)

Although effectiveness has been proven, the practical implementation process also revealed significant challenges, correctly reflecting concerns about the gap between technology and science teacher competency. The biggest challenge lies in the Evaluate step, where students felt confused when detecting subtle scientific errors generated by AI. This confirms the dialectical relationship in the TPACK framework: technology cannot compensate for a lack of specialized science content knowledge. If students do not master Science PCK, they cannot effectively perform the control role. This finding aligns with research suggesting that scientific verification competency needs to be considered a new core professional competency for science teachers (Baxmann et al., 2025; Lindell & Utterberg Modén, 2025). The second challenge relates to time pressure and the psychological hesitation to revise during the Iterate step. Having to continuously refine scientific prompts and products requires higher patience and cognitive effort, which is a necessary element to prevent pedagogical shallowness in science instruction. Additionally, issues of algorithmic bias also appeared sporadically, requiring students to have high ethical sensitivity for screening, suggesting that fostering AI competency for science teachers cannot be separated from AI ethics education (Barnes & Tour, n.d.; Pandey et al., 2025).

5.4. Limitations of the Study

Although the study achieved positive results in constructing and verifying the AI-REFINE process, certain limitations exist that need careful consideration when interpreting findings.

First, regarding the expert panel in Phase 1, while the sample size of 09 experts aligns with the standard requirements for the Delphi method to reach consensus (Polit & Beck, 2006), a limitation lies in the demographic scope of the panel. All participating experts are from Vietnam. The absence of international experts may limit the diversity of perspectives regarding global AI education standards and the cross-cultural applicability of the process. Future studies would benefit from incorporating a more international panel to enhance the global validity of the model.

Second, regarding the experimental design in Phase 2, the study was conducted on a small group of pre-service natural science teachers (N = 28) majoring in Natural Sciences at a single training institution. This may limit the representativeness and the ability to generalize results to other educational cultural contexts or Social Sciences majors. Furthermore, the use of a one-group pre-test-post-test design without a randomized control group means the study could not fully control for potential extraneous variables that might influence experimental results.

Finally, the intervention duration of 06 weeks, while sufficient to record initial shifts in competency, lacks sufficient data to confirm the long-term sustainability of co-design skills in students after the training process ends, suggesting a need for longitudinal studies in the future.

6. Conclusion and Recommendations

6.1. Conclusion

This study achieved its primary objective of developing and content-validating the AI-REFINE process to foster the science-integrated STEM lesson plan co-design competency of pre-service science teachers. Through a rigorous two-round Delphi technique involving science education experts, the six-step R-E-F-I-N-E model was established with an S-CVI/Ave = 0.97. Results from the pilot study involving 28 pre-service teachers demonstrated the practical feasibility and promising potential of the process when applied in authentic teacher education contexts. Quantitative analysis indicated a statistically significant improvement in lesson plan design competency scores following the intervention, with the mean score increasing from 5.77 to 7.80. The large effect size (Cohen's d_{av} = 1.80) highlights the process's capacity to effectively support the enhancement of students' design skills, even within a small sample size. The AI-REFINE process guides students to transition from passive AI users to active co-designers, helping them maintain pedagogical control and ensure scientific accuracy through the human-in-the-loop philosophy. Although the initial findings are highly encouraging, this study acknowledges the limitation of lacking a control group in its design, thereby laying the groundwork and suggesting directions for larger-scale experimental research in the future.

6.2. Recommendations

This study recommends that science teacher training institutions consider integrating the AI-REFINE process into formal curricula, particularly within specialized teaching methodology or educational technology application modules. For effective implementation, priority should be given to the professional development of teacher educators, ensuring they possess adequate digital competency and specific scientific critical thinking to accurately model and guide students through the process. Furthermore, future research directions should expand the scope of investigation through replication studies across various scientific disciplines and multiple training institutions to verify the model's generalizability. Subsequent studies should also employ rigorous experimental designs with larger sample sizes and randomized control groups, coupled with longitudinal approaches to evaluate the long-term sustainability of the co-design competency post-intervention, thereby contributing to the comprehensive theoretical framework of artificial intelligence application in science teacher education.

Declarations

Ethics approval

This study involving human participants was reviewed and approved by the Ethics Committee of the University of Education - Thai Nguyen University. The research procedures adhered to the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to participate

Informed consent was obtained from all experts participating in the Delphi study and all pre-service teachers involved in the pedagogical experiment

Clinical trial number

Clinical trial number: not applicable.

Consent for publication

Informed consent was obtained from all individual participants included in the study for the publication of anonymized data.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to participant privacy protection but are available from the corresponding author, Nguyen Quang Linh, on reasonable request via email: linhnq@tnue.edu.vn.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Authors' contributions

All authors contributed to the study conception and design. Data collection was performed by Do Huong Tra and Nguyen Quang Linh. Data analysis and processing were performed by Nguyen Quang Linh. The first draft of the manuscript was written by Do Huong Tra, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors wish to thank the panel of nine experts for their valuable contribution and consensus in the Delphi study. We also extend our gratitude to the pre-service teachers at the University of Education - Thai Nguyen University for their enthusiastic participation in the pedagogical experiment.

References

1. Abate A, Atnafu M, Michael K. Visualization and problem-based learning approaches and students' attitude toward learning mathematics. Pedagogical Res. 2022;7(2):1–13. https://doi.org/10.29333/pr/11725.

2. Abbonizio J, Palermo C, Brand G, Buus N, Fossey E, Dart J. Co-designing formal health professions curriculum in partnership with students: A scoping review. Med Teach. 2025;47(3):413–24. https://doi.org/10.1080/0142159X.2024.2339403.
3. Al-Nuaimi SR, Al-Ghamdi SG. (2022). Assessment of knowledge, attitude and practice towards sustainability aspects among higher education students in Qatar. *Sustainability*, *14*. https://doi.org/10.3390/su142013149
4. Alrwaished N. Mathematics pre-service teachers' preparation program for designing STEM based lesson plan: Enhanced skills and challenges. Cogent Educ. 2024;11(1):2320467. https://doi.org/10.1080/2331186X.2024.2320467.
5. Alsaleh NJ. (2020). Teaching critical thinking skills: Literature review. *Turkish Online Journal of Educational Technology-TOJET*, *19*. https://files.eric.ed.gov/fulltext/EJ1239945.pdf
6. Azcárate ALV. (2024). Foresight methodologies in responsible GenAI education: Insights from the Intermedia-Lab at Complutense University Madrid. *Education Sciences*, *14*. https://doi.org/10.3390/educsci14080834
7. Ba S, Hu X, Stein D, Liu Q. Assessing cognitive presence in online inquiry-based discussion through text classification and epistemic network analysis. Br J Edu Technol. 2023;54. https://doi.org/10.1111/bjet.13285.
8. Barnes M, Tour E. (n.d.). Teachers' use of generative AI: a 'dirty little secret'? Lang Educ, *40*(1), 24–39. https://doi.org/10.1080/09500782.2025.2485935
9. Baxmann M, Kárpáti K, Baráth Z. The potentials and challenges of integrating generative artificial intelligence (AI) in dental and orthodontic education: A systematic review. BMC Oral Health. 2025;25(1):905. https://doi.org/10.1186/s12903-025-06070-7.
10. Denzin NK, Lincoln YS. The Sage handbook of qualitative research. 4th ed. SAGE; 2018. https://us.sagepub.com/en-us/nam/the-sage-handbook-of-qualitative-research/book242504.
11. Đỗ TT. Ứng dụng ChatGPT trong giáo dục đại học: Cơ hội, thách thức và một số khuyến nghị [Application of ChatGPT in higher education: Opportunities, challenges, and recommendations]. Tạp Chí Giáo Dục. 2025;25(5):453–9. https://tcgd.tapchigiaoduc.edu.vn/index.php/tapchi/article/view/3937.
12. Fenta AA. A review on enhancing education with AI: Exploring the potential of ChatGPT, Bard, and generative AI. Discover Educ. 2025;4(1):38. https://doi.org/10.1007/s44217-025-00426-5.
13. Ganesan U, Morales AR. A science teacher's experiences when fostering intercultural competence among students in multilingual classrooms: A narrative study. Cult Sci Edu. 2024;19(1):189–208. https://doi.org/10.1007/s11422-023-10206-y.
14. García-Carmona A. Scientific thinking and critical thinking in science education. Sci Educ. 2025;34(1):227–45. https://doi.org/10.1007/s11191-023-00460-5.
15. Herro D, Adisa IO, Abimbade O. Teachers co-designing and enacting elementary data science curriculum through connected learning. J Stat Data Sci Educ. 2025;00(0):1–19. https://doi.org/10.1080/26939169.2025.2459217.

16. Holmes W, Miao F. Guidance for generative AI in education and research. Unesco Publishing. 2023. https://doi.org/10.54675/KMQF4065.
17. Jang S-J, Chen K-C. From PCK to TPACK: Developing a transformative model for pre-service science teachers. J Sci Edu Technol. 2010;19(6):553–64. https://doi.org/10.1007/s10956-010-9222-y.
18. Kılıçkaya F, Kic-Drgas J. Pre-service language teachers' experiences and perceptions of integrating generative AI in practicum-based lesson study. Humanit Social Sci Commun. 2025;12(1):1478. https://doi.org/10.1057/s41599-025-05715-w.
19. Kurt G, Çakıroğlu E. Prospective mathematics teachers' TPACK development in graphical data displays: A microteaching lesson study. Int J Math Educ Sci Technol. 2025;56(3):445–75. https://doi.org/10.1080/0020739X.2023.2246970.
20. Lê VMT, Nguyễn TTA. Đề xuất quy trình thiết kế kế hoạch bài dạy môn Toán ở tiểu học dựa trên khung kiến thức toán học sư phạm công nghệ (TPMK) [Proposing a process for designing primary mathematics lesson plans based on the technological pedagogical mathematical knowledge (TPMK) framework]. Tạp Chí Giáo Dục. 2025;25(14):30–5. https://tcgd.tapchigiaoduc.edu.vn/index.php/tapchi/article/view/3660.
21. Lindell TL, Utterberg Modén M. Conflicting motives: Challenges of generative AI in education. Learn Media Technol. 2025;50(3):324–37. https://doi.org/10.1080/17439884.2024.2438933.
22. Liu J, Wang C, Liu Z, Gao M, Xu Y, Chen J, Cheng Y. A bibliometric analysis of generative AI in education: Current status and development. Asia Pac J Educ. 2024;44(1):156–75. https://doi.org/10.1080/02188791.2024.2305170.
23. Mahat M, Dollinger M, D'Angelo B, Naylor R, Harvey A. Co-designing a curriculum model for career education: Perspectives from regional communities in Australia. Australian Educational Researcher. 2023;50(2):409–31. https://doi.org/10.1007/s13384-021-00505-0.
24. Mejeh M, Sarbach L. Co-design: From understanding to prototyping an adaptive learning technology to enhance self-regulated learning. Technol Knowl Learn. 2025;30(2):1021–46. https://doi.org/10.1007/s10758-024-09788-5.
25. Misawa T, Koizumi A, Tamura R, Yoshimi K. Exploring utilization of generative AI for research and education in data-driven materials science. Sci Technol Adv Materials: Methods. 2025;5(1):2535956. https://doi.org/10.1080/27660400.2025.2535956.
26. Mustafy T, Rahman MTU. (2024). SPSS. In T. Mustafy & Md. T. U. Rahman, editors, *Statistics and data analysis for engineers and scientists* (pp. 135–180). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-4661-7_4
27. Nguyen KV. The use of generative AI tools in higher education: Ethical and pedagogical principles. J Acad Ethics. 2025;23(3):1435–55. https://doi.org/10.1007/s10805-025-09607-1.
28. Nguyễn QL, Vũ TTY, Dương VT, Nông M, Ấn.. Những năng lực giáo viên cần được bồi dưỡng để triển khai thành công dạy học theo định hướng giáo dục STEM: nghiên cứu trường hợp tại tỉnh Thái Nguyên [The competencies teachers need to develop for successful implementation of STEM-

oriented teaching: A case study in Thai Nguyen province]. Tạp Chí Giáo Dục. 2023;23(6):51–7. https://tcgd.tapchigiaoduc.edu.vn/index.php/tapchi/article/view/676.

29. Pandey CS, Mishra P, Pandey SR, Pandey S. (2025). Epistemic trust in generative AI for higher education scale (ETGAI-HE scale). *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02566-6
30. Penuel WR, Allen A-R, Henson K, Campanella M, Patton R, Rademaker K, Reed W, Watkins D, Wingert K, Reiser B, Zivic A. Learning practical design knowledge through co-designing storyline science curriculum units. Cognition Instruction. 2022;40(1):148–70. https://doi.org/10.1080/07370008.2021.2010207.
31. Phạm KC, Lê CN, Hoàng BL. Đề xuất mô hình đồng sáng tạo cùng AI và vận dụng trong thiết kế hoạt động dạy học: Nghiên cứu trường hợp dạy học lập trình với sự hỗ trợ của Chatbot AI cho sinh viên đại học [Proposing an AI co-creation model and its application in designing teaching activities: A case study of teaching programming with AI chatbot support for university students]. Tạp Chí Giáo Dục. 2025;25(16):54–9. https://tcgd.tapchigiaoduc.edu.vn/index.php/tapchi/article/view/3843.
32. Polit DF, Beck CT. The content validity index: Are you sure you know what's being reported? critique and recommendations. Res Nurs Health. 2006;29(5):489–97. https://doi.org/10.1002/nur.20147.
33. Qian Y. Pedagogical applications of generative AI in higher education: A systematic review of the field. TechTrends. 2025;69(5):1105–20. https://doi.org/10.1007/s11528-025-01100-1.
34. Schiering D, Sorge S, Keller MM, Neumann K. A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK)—What constitutes high-level PCK? J Res Sci Teach. 2023;60(1):136–63. https://doi.org/10.1002/tea.21793.
35. Shulman LS. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*. https://doi.org/10.17763/haer.57.1.j463w79r56455411
36. Smith A, McConnell L, Iyer P, Allman-Farinelli M, Chen J. Co-designing assessment tasks with students in tertiary education: A scoping review of the literature. Assess Evaluation High Educ. 2025;50(2):199–218. https://doi.org/10.1080/02602938.2024.2376648.
37. Wang X, Niu J, Fang B, Han G, He J. Empowering teachers' professional development with LLMs: An empirical study of developing teachers' competency for instructional design in blended learning. Teach Teacher Educ. 2025;165:105091–105091. https://doi.org/10.1016/j.tate.2025.105091.
38. Wardak D, Wilson S, Zeivots S. Co-design as a networked approach to designing educational futures. Postdigital Sci Educ. 2024;6(1):194–210. https://doi.org/10.1007/s42438-023-00425-5.
39. Wiggins GP, McTighe J. Understanding by design. ASCD; 2005.
40. Wu C, Carroll JM. Evolving perspectives on generative AI in higher education: Insights from two years of university stakeholder analysis. TechTrends. 2025. https://doi.org/10.1007/s11528-025-01150-5.
41. York RO. (2022). Using SPSS to analyze your outcome data. In R. O. York, editor, *Evaluating Human Service Outcomes* (pp. 135–159). Springer International Publishing. https://doi.org/10.1007/978-3-031-10175-5_6

42. Zeivots S, Hopwood N, Wardak D, Cram A. Co-design practice in higher education: Practice theory insights into collaborative curriculum development. High Educ Res Dev. 2025;44(3):769–83. https://doi.org/10.1080/07294360.2024.2410269.

Figures

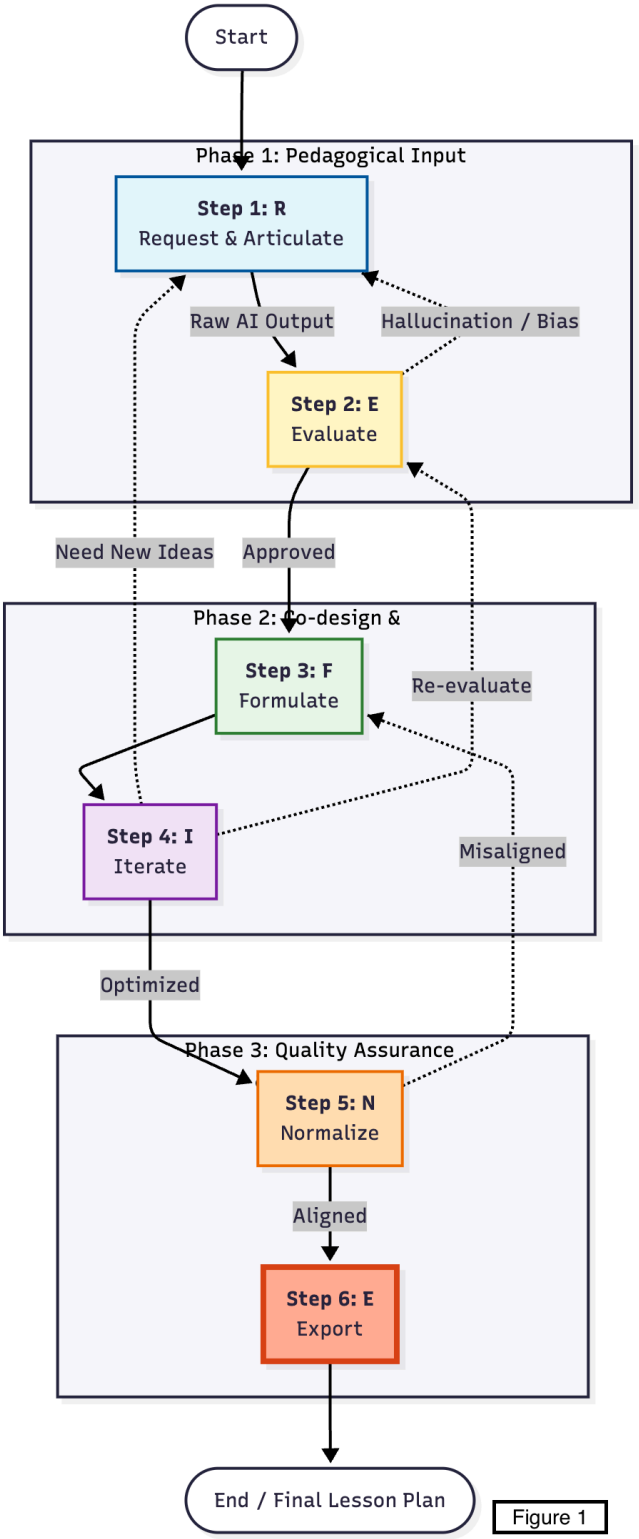


Figure 1

The AI-REFINE Process Model: A six-step framework for fostering PSTs' co-design competency with Generative AI

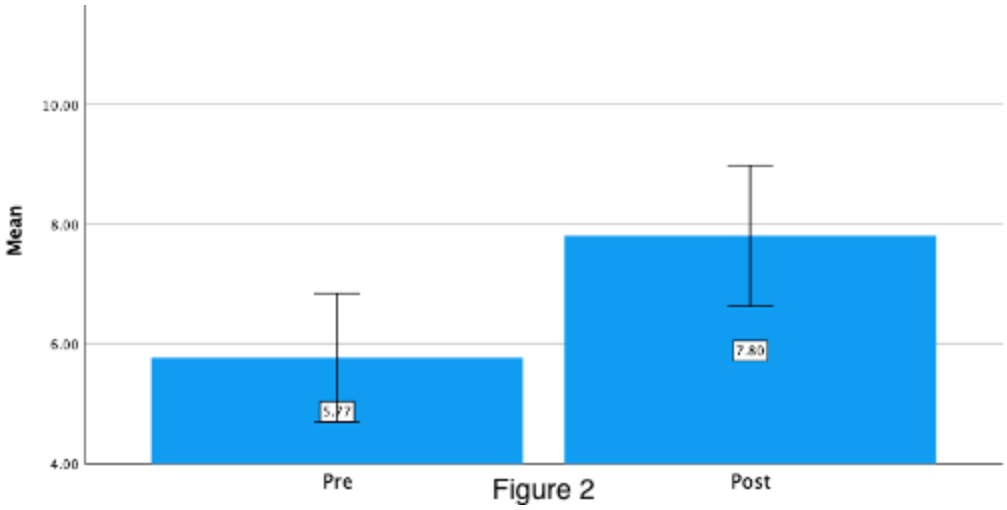


Figure 2

Comparison of pre-service teachers’ co-design competency scores before and after the intervention.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SurveyQuestionnaireRound1Delphi.docx
- SurveyQuestionnaireRound2Delphi.docx
- DatabySPSS26.0.docx