frontiers in Education

TYPE Original Research
PUBLISHED 18 June 2026
DOI 10.3389/feduc.2026.1859688

OPEN ACCESS

EDITED BY
Doniwen Pietersen,
University of South Africa, South Africa

REVIEWED BY
Irfan Ahmed Rind,
Sukkur IBA University, Pakistan
Wiets Botes,
Sol Plaatje University, South Africa

*CORRESPONDENCE
Naiara Bilbao-Quintana
naiara.bilbao@ehu.eus

RECEIVED 19 April 2026
REVISED 27 May 2026
ACCEPTED 31 May 2026
PUBLISHED 18 June 2026

CITATION
Bilbao-Quintana N, Tejada Garitano E, Romero Andonegui A and López de la Serna A (2026) From AI assistance to pedagogical reflection: a rubric-mediated model for generative AI in teacher education.
Front. Educ. 11:1859688.
doi: 10.3389/feduc.2026.1859688

From AI assistance to pedagogical reflection: a rubric-mediated model for generative AI in teacher education

Naiara Bilbao-Quintana*, Eneko Tejada Garitano, Ainara Romero Andonegui and Arantzazu López de la Serna

Department of Didactics and School Organization, Faculty of Education, University of the Basque Country (EHU), Leioa, Spain

The integration of Generative Artificial Intelligence (GenAI) into teacher education necessitates a transition from technical proficiency to critical pedagogical judgment. This study adopts a Design-Based Research (DBR) approach to evaluate the impact of an analytical mediation model based on cognitive oversight rubrics. Utilizing a sample of pre-service teachers ($N = 195$), the study analyzed the evolution of perceived self-efficacy and the quality of didactic artifacts designed using Large Language Models (LLMs). Resultsindicate significant improvement across all dimensions, notably the strengthening of pedagogical judgment ($t = 35.12$; $p < .001$) with a large effect size ($d > 0.80$). A pivotal finding is the emergence of the self-efficacy paradox: while actual design competence and quantitative self-efficacy significantly increased, qualitative evidence demonstrates that this confidence became more cautious, reflective, and critically calibrated, indicating a shift toward conscious, self-regulated competence. However, persistent variability in STEAM complexity suggests that technological scaffolding encounters a performance ceiling when lacking a solid foundation of Pedagogical Content Knowledge (PCK). The study concludes that human-in-the-loop design and critical mediation are essential to mitigate cognitive offloading and ensure teacher epistemic agency in the algorithmic era.

KEYWORDS
design-based research, evaluative judgement, generative AI, human-centered AI, self-regulated learning, teacher education

1 Introduction

The emergence of Generative Artificial Intelligence (GenAI) in Higher Education has transitioned, with remarkable speed, from an initial moral panic centered on plagiarism toward a pragmatic acceptance that often lacks pedagogical depth (Kasneci et al., 2023; Luckin, 2024; UNESCO, 2023). Moving past the reactive phase, academic debate now focuses on the ethical governance of AI and its integration into instructional design, demanding frameworks that ensure teacher intellectual sovereignty (Dwivedi et al., 2023; Shneiderman, 2022; Zawacki-Richter et al., 2019). However, a critical gap persists: while students develop operational technical competence, epistemic agency and professional judgment are diluted in automation processes that prioritize the product over pedagogical reflection (Caspari-Sadeghi, 2026).

The fundamental problem is not the capacity of Large Language Models (LLMs) to generate knowledge, but the erosion of teacher authority that occurs when decision-making is delegated to algorithms without a critical mediation framework (Bearman and Ajjawi, 2023; Floridi, 2023). Recent literature warns that high perceived self-efficacy does not necessarily correlate with actual competence for rigorous didactic integration (Kruger and Dunning, 1999; Zimmerman, 2000). This illusion of competence poses a systemic risk in initial teacher education: the creation of educational designs that are formally correct but lack situated semantics and a solid ethical foundation (Chaudhry et al., 2023; Molenaar, 2022). Ensuring this agency is not merely a technical requirement but a cornerstone for human flourishing in education, as it prevents the reduction of teaching to a mechanistic process, preserving the educator's role as a moral and social guide. In this light, the present research asks: How can an analytical mediation artifact transform the interaction with GenAI from passive cognitive offloading into an exercise of pedagogical oversight?

Within this scenario, mediation cannot be reduced to a mere user guide; it must be framed as a socio-cognitive artifact that compels the user to reclaim design control and epistemic responsibility. This study proceeds from the premise that rubrics—traditionally regarded as summative assessment instruments—can be reconfigured as cognitive oversight tools and analytical scaffolding within AI-mediated design. By imposing explicit pedagogical quality criteria and constraints, the rubric serves as a deliberative brake, enabling students to navigate the generative overabundance of LLMs without compromising their evaluative judgment. In this sense, reclaiming evaluative judgment from algorithmic opacity is a democratic necessity; it empowers future teachers to act as critical citizens who can navigate and challenge automated systems rather than being passively governed by them.

The objective of this research is to evaluate the impact of an intervention model based on Design-Based Research (DBR), which utilizes cognitive oversight rubrics to transition from a purely technical interaction with AI toward professional pedagogical agency. By addressing the risks of automated passivity, this research contributes to SDG 4 (Quality Education) by developing frameworks that safeguard the epistemic integrity of future teachers. Through an analysis of perceived self-efficacy and the epistemic quality of artifacts designed by $N = 195$ pre-service teachers across various academic levels, the study examines how analytical mediation helps mitigate the self-efficacy paradox, transforming potential metacognitive passivity into self-regulated learning (SRL) processes. The study analyzes whether this critical scaffolding shifts the focus from the AI-generated product toward reflective practice, structured around the following research questions (RQ):

- RQ1. Impact on Design Quality: To what extent does mediation through cognitive oversight rubrics shift the quality of GenAI-designed artifacts from technical adequacy toward professional pedagogical proficiency?
- RQ2. The Self-Efficacy Paradox: How does the mediated critical evaluation process affect pre-service teachers' perception of self-efficacy compared to their objective performance in instructional design?
- RQ3. The Performance Ceiling: Are there significant differences in the progression of technical and pedagogical dimensions, and what role does Pedagogical Content Knowledge (PCK) play in overcoming the performance plateau of large language models?

2 Theoretical framework

2.1 Towards a human-centered AI (HCAI) model in teacher education

The Human-Centered Artificial Intelligence (HCAI) paradigm posits that technology should function as an amplifier of human capability rather than a substitute, ensuring that users retain sovereign control through a Human-in-the-loop approach (Sahlberg, 2024; Shneiderman, 2022). In the field of teacher education, this perspective requires GenAI to be integrated as a cognitive partner whose pedagogical effectiveness is strictly contingent upon the robustness of the instructional design. This vision is supported by prior research positioning technology as a vector for metacognitive development. For instance, recent studies by the team in hybrid environments suggest that such mediation can improve self-regulated learning (SRL) by up to 22%. Consequently, this approach allows higher-order executive functions to become visible and conscious for the student (Yañez-Perea et al., 2025a). In this regard, recent systematic reviews (Li and Li, 2026) confirm that the development of critical and creative thinking through language models depends on the presence of explicit scaffolds. In the absence of such mediation, students tend to prioritize algorithmic immediacy, resulting in cognitive offloading processes that neutralize deep reflection and professional decision-making (Castañeda and Williamson, 2021). In this sense, HCAI in teacher education is not merely a technical preference but a democratic safeguard. By ensuring that the educator remains "in the loop," we prevent the delegation of ethical and social responsibilities to opaque systems, thereby fostering a pedagogical environment conducive to human flourishing and critical citizenship.

In this scenario, the Technological Pedagogical Content Knowledge (TPACK) framework (Mishra and Koehler, 2006) assumes renewed relevance, yet it also faces new risks of erosion. While GenAI expands technological affordances, its uncritical use can lead to concerning metacognitive passivity and intellectual dependence that displaces the productive struggle necessary for professional learning (Selwyn, 2024; Tao et al., 2026). Faced with this inertia, it is imperative to design scenarios that promote an explicit "culture of thinking", similar to that articulated in the *Pentsatu* conceptual space (Yañez-Perea et al., 2025b). Under this paradigm, interaction with AI must compel the student toward "cognitive homeostasis"—that is, a self-regulation process in which the subject reconfigures their knowledge in response to the algorithm's speed to avoid techno-dependency. Without a solid foundation of PCK (Pedagogical Content Knowledge), AI resources reduce the teacher's role to a mere syntactic validator of content, transforming didactic design into a blind delegation that dilutes pedagogical authority and curricular integrity.

In this algorithmic landscape, the introduction of Generative AI does not merely provide technical assistance; it fundamentally redistributes the teacher's cognitive labor. As Rind (2026) argues, teachers must be conceptualized as "cognitive orchestrators" who require active pedagogical oversight to maintain professional control over automated processes. When AI systems accelerate instructional production, they threaten to displace the deeply reflective and diagnostic practices through which authentic teacher expertise and Pedagogical Content Knowledge (PCK) are traditionally developed (Rind and Qureshi, 2026). Without explicit scaffolds, uncritical reliance on large language models risks causing cognitive offloading, where the technical fluency of the machine silences the evaluative judgment of the educator. Therefore, human-centered AI integration requires intentional mechanisms designed not to grade the final product, but to preserve human agency and guarantee that technology amplifies, rather than replaces, the critical deliberative labor of teaching.

Consequently, the cognitive oversight rubric presented in this study is proposed as an empirical response to the need for AI-mediated pedagogy. This artifact should not be understood as a simple assessment instrument, but rather as the necessary interface to ensure that pre-service teachers maintain their epistemic agency. To guarantee this purpose, the rubric's design follows expert judgment validation protocols similar to those successfully employed in the development of instruments for the IkasLab project, where scientific rigor was ensured through the Delphi method and an expert competence coefficient of $K = 0.835$ (Yañez-Perea et al., 2025b). By imposing constraints and analytical quality criteria, the rubric acts as a deliberative brake that compels students to interrogate the algorithm based on sound didactic foundations, transforming a potentially passive interaction into an exercise in rigorous and professionally grounded pedagogical engineering.

2.2 Self-regulated learning (SRL) and the illusion of competence

One of the greatest challenges identified in recent literature is the gap between perceived self-efficacy and actual competence in the use of LLMs. According to Self-Regulated Learning (SRL) theory, success in complex tasks depends on the subject's ability to monitor and evaluate their own cognitive processes (Zimmerman, 2000). However, the ease of use and eloquence of Generative AI can induce a process of blind cognitive offloading, where students relax their metacognitive vigilance, perceiving that the machine has already solved the design problem.

This phenomenon is directly linked to the Dunning-Kruger effect: individuals with lower actual competence tend to overestimate their abilities due to their inability to recognize the task's complexity (Kruger and Dunning, 1999). In teacher education, this manifests as an illusion of competence: pre-service teachers believe they have mastered educational design because the AI generates a formally correct structure, while remaining oblivious to biases, hallucinations, and a lack of deep pedagogical coherence. Therefore, developing critical evaluative judgment becomes imperative to transform AI usage from passive cognitive discharge into reflective scaffolding.

2.3 Analytical mediation, scaffolding, and cognitive load

From a socio-constructivist perspective, mediation is the fundamental mechanism for developing higher psychological functions. The analytical mediation rubric proposed in this study acts as a scaffolding artifact (Vygotsky, 1978) that redistributes the subject's cognitive load. According to Cognitive Load Theory (Sweller, 1988), AI-assisted educational design imposes a high intrinsic load due to the need to simultaneously process the logic of the tool and the logic of the curriculum.

The rubric functions as a deliberative brake that reduces extraneous load and directs student attention toward critical design aspects (objectives, ethics, interdisciplinarity). This scaffolding is particularly necessary in high-complexity environments such as the STEAM approach, where interdisciplinary integration requires a synthesis that AI often addresses in a fragmentary manner. Analytical mediation ensures that the human-AI interaction remains within the Zone of Proximal Development (ZPD), forcing students to exercise constant oversight and preventing automation from neutralizing the cognitive effort required for deep learning. Furthermore, this analytical scaffolding plays a crucial role in promoting educational equity (SDG 10). By providing explicit criteria for oversight, the rubric levels the playing field, ensuring that all pre-service teachers—regardless of their prior technological capital—can exercise high-level pedagogical agency and avoid the risks of algorithmic bias or exclusion.

2.4 Design-based research (DBR) as a transformative paradigm

To investigate these processes, this study is framed within Design-Based Research (DBR). DBR moves away from decontextualized experimental methods to focus on the development and refinement of educational interventions in real-world settings (McKenney and Reeves, 2018). This paradigm allows the pedagogical intervention to be treated not as an isolated variable, but as a complex system where the tensions between the tool, the scaffolding, and the subject are analyzed. The ultimate goal of DBR is not merely to validate the rubric's efficacy, but to generate design principles that can be transferred to other higher education contexts, ensuring an AI integration that is ethical, critical, and pedagogically grounded.

2.5 Operationalization and empirical indicators of core constructs

To ensure conceptual precision and support the mixed-methods alignment of this study, the key theoretical constructs are operationally defined and linked to their specific empirical indicators as follows:

- Pedagogical Judgment: Grounded in the work of Rind (2026), it is defined as the teacher's capacity to deploy professional expertise, manage redistributed cognitive load, and make deliberate curricular decisions under AI-

mediated conditions. In this study, this construct was measured quantitatively through a dedicated self-report subscale within the Perceived Self-Efficacy Questionnaire (capturing students' perceived capability to make sound instructional choices, yielding the reported $t = 35.12$ shift) and qualitatively triangulated via the reflective narrative corpus.

- Epistemic Agency: Drawing on the continuum proposed by Rind et al. (2026), this construct represents the student's active cognitive autonomy and progression away from *blind trust* in automated LLM outputs toward *critical inquiry, systematic verification, and independent evaluation*. Rather than a static trait, it is operationalized as the quality of students' active interrogation during the prompt-looping process, and was measured qualitatively through the thematic analysis of the mandatory open-ended reflection fields in Phase 3.
- Evaluative Judgment: Defined as the capacity to critically assess the quality of an educational design (regardless of whether it was generated by a human or an AI) against explicit, objective performance standards. In this research, this construct was measured objectively and empirically through the pre/post evolution of the lesson plan quality scores evaluated via the 5-dimension Analytical Rubric (as detailed in Tables 1, 2).
- Algorithmic Sovereignty: Defined as the human-in-the-loop governance mechanism through which the educator maintains ultimate pedagogical authority, preventing uncritical cognitive offloading and automated complacency. Empirically, it is manifested through the practical execution of the rubric as a metacognitive deliberative brake that forces the reconfiguration of generic AI text into situated curricular scenarios.

3 Methodology

3.1 Research design: DBR approach and HCAI design principles

A Design-Based Research (DBR) approach (McKenney and Reeves, 2018) was adopted, providing an ideal framework for analyzing AI integration through the lens of Human-Centered Artificial Intelligence (HCAI). This methodological choice aligns with contemporary trends investigating the reliability of large language models in analyzing the quality of lesson plans in initial teacher education (Hauk and Soujon, 2026). Similar to the CODE-PLAN framework, which focuses on content transformation and sequencing, DBR enables the addressing of complex educational problems through the creation of cognitive scaffolding artifacts in real-world contexts. In this study, the artifact is not the GenAI technology *per se*, but rather the analytical mediation rubric designed to expand students' epistemic agency and ensure a level of didactic rigor that pure automation cannot guarantee.

The intervention design is aligned with the pedagogical and technological principles developed within a broader international initiative, specifically the Horizon Europe STEAMbrace project, with a focus on the integration of STEAM approaches and AI-supported learning environments. The rubric-mediated model implemented in this study can be understood as a micro-level instantiation of these broader design principles, focusing on the development of teacher agency and critical interaction with generative AI systems.

To fully operationalize the iterative and evolutionary nature of the DBR approach, the intervention was structured around one overarching macro-cycle containing two operational micro-iterative sub-cycles, moving away from a traditional linear pre-post design.

Micro-cycle 1 (Prompt-Output Iteration): Centered on the student-AI interaction loop. Empirical evidence from initial prompt outcomes revealed a high density of pedagogical hallucinations and superficial compliance with the Basque curriculum. This evidence forced an immediate collaborative redesign of students' prompting strategies, moving from single-turn inputs to iterative, prompt-looping sequences guided by the emerging criteria.

Micro-cycle 2 (Artifact Refinement): Centered on the mediation tool itself. The initial draft of the analytical rubric underwent a formative evaluation based on user-experience feedback and inter-rater reliability analysis during the first two weeks of Phase 3. Qualitative evidence of criteria ambiguity regarding interdisciplinary synthesis led to the physical redesign of the 'STEAM Complexity' descriptors. These descriptors were sharpened with explicit operational examples to reduce cognitive overload and maximize the rubric's effectiveness as a metacognitive scaffold.

Through these recursive evaluation and refinement loops, the final high-level design principles for Human-Centered AI (HCAI) integration in teacher training were synthesized, ensuring that the final artifact was directly shaped by empirical field evidence.

3.2 Context and participants

The research was conducted with a sample of $N = 195$ students from the Bachelor's Degree in Primary Education at a public university in the Basque Country, Spain. The sample was stratified by academic level, including first-year ($n = 106$) and second-year ($n = 89$) students enrolled in two courses associated with the Department of Didactics and School Organization. A non-probabilistic convenience sampling method was employed, which is consistent with the iterative refinement cycles inherent to Design-Based Research (DBR). Although participants reported regular use of digital technologies, they lacked prior instruction regarding the affordances and risks of Large Language Models (LLMs) in curricular design. To ensure the validity of the aggregate analysis, both groups received the same intervention protocol, mediation rubrics, and GenAI tools.

Ethical Considerations: This study was conducted in accordance with European data protection regulations (GDPR, Regulation (EU) 2016/679) and institutional guidelines for educational research. Participation was voluntary and based on informed consent; all participants were informed of the study's purpose, data usage, and their right to withdraw at any time without academic consequences. Regarding data integrity, because the digital instruments deployed on the institution's platform utilized mandatory submission validation fields,

TABLE 1 Descriptive statistics of pedagogical design quality across the three phases ($N = 195$).

Dimension	Human-only design (M/*SD*)	Autonomous AI design (*M*/*SD*)	Rubric-mediated AI design (*M*/*SD*)
Pedagogical coherence	2.12/0.78	3.24/0.89	4.31/0.52
Curricular specification	2.35/0.64	3.41/0.72	4.15/0.58
Competency-based objectives	1.98/0.82	3.85/0.94	4.48/0.41
Learning situation design	2.21/0.88	3.08/0.96	4.12/0.69
Interdisciplinary/STEAM integration	1.84/0.91	2.95/1.15	3.82/0.92

participants could not leave items blank. Consequently, there were no missing data or dropped items (0% missingness/attrition), allowing all statistical procedures to be executed on the full sample of $N = 195$ complete cases.

Data were collected and processed anonymously, ensuring that no personally identifiable information (PII) was recorded or stored. Given the nature of the study—focused on routine educational practice within a higher education course—and the absence of sensitive personal data, formal approval from an institutional ethics committee was not required according to local regulations. The research design also adhered to the ethical framework established within the broader international research initiative Horizon Europe STEAMbrace under which this study is framed, including internal procedures to guarantee Responsible Research and Innovation (RRI) principles.

3.3 Instruments and intervention artifacts

The study is centered on two *ad hoc* instruments:

1. Perceived Self-Efficacy Questionnaire (Pre/Post): A 12-item scale (Likert 1-5) with high overall internal consistency (Cronbach's $\alpha = .88$). To ensure psychometric rigor, reliability was analyzed for each individual subscale: the operational prompt engineering competence subscale yielded a Cronbach's $\alpha = .82$; the curricular integration subscale achieved an $\alpha = .85$; and the ability to detect algorithmic bias subscale demonstrated an $\alpha = .80$. It measures confidence across these three dimensions.
2. Analytical Scaffolding Artifact (The Rubric): Grounded in the principles of Cognitive Oversight, this tool functions as a deliberative brake on AI affordances. Moving beyond the traditional view of a rubric as a static grading sheet, this artifact is operationally conceptualized as a cognitive rebalancing mechanism (Rind, 2026; Rind and Qureshi, 2026). Its primary design objective is to safeguard teacher agency against the displacement of reflective planning practices. It evaluates three axes:

    - Syntactic Dimension: Structural coherence of the output.
    - Situated Semantic Dimension: Degree of content transfer and adaptation to the real-world context (as opposed to generic AI-generated output).
    - STEAM Complexity Dimension: Quality of the interdisciplinary synthesis.

TABLE 2 Repeated measures ANOVA results for design dimensions ($N = 195$).

Dimension	F (2,388)	p	η_p^2
Pedagogical coherence	342.15	< 0.001	0.64
Curricular specification	215.48	< 0.001	0.52
Competency-based objectives	401.82	< 0.001	0.67
Learning situation design	189.33	< 0.001	0.49
Interdisciplinary/STEAM integration	94.67	< 0.001	0.33

To capture the qualitative dimension of students' epistemic agency and complement the quantitative metrics, the digital instruments administered during the evaluation phase included mandatory open-ended justification fields. In these qualitative text boxes, participants were prompted to critically reflect on the discrepancies between their autonomous AI designs and their final rubric-mediated refinements, as well as the immediate shifts in their perceived teaching competence. This embedded technique generated a qualitative corpus of narrative evidence from the 195 participants, which was subsequently anonymized using alphanumeric codes (e.g., Subject 84, Subject 159) and subjected to thematic analysis to contextualize the statistical transitions.

3.4 Procedure and implementation

The instructional task assigned to the pre-service teachers consisted of designing a comprehensive lesson plan (locally operationalized as a "Learning Situation" under the current educational curriculum). In this study, the 'STEAM approach' was adopted as an integrated instructional design framework rather than a mere thematic topic. This meant that the lesson planning process was strictly bound to the principles of interdisciplinary integration, requiring students to generate instructional sequences where Science, Technology, Engineering, Arts, and Mathematics synergistically converged to solve an authentic, real-world problem. Consequently, the STEAM framework directly conditioned the lesson planning by demanding active methodologies—such as Problem-Based Learning (PBL)—and specific cross-disciplinary assessment metrics, which were subsequently evaluated through the oversight rubric.

The intervention was deployed over six weeks, following four iterative phases inherent to DBR:

- Phase 1: Diagnosis. Administration of the pre-test to establish a self-efficacy baseline.

- Phase 2: Autonomous Design and Exploration. Students generated a learning situation through the unrestricted use of LLMs (ChatGPT, Claude, or Gemini). This phase was characterized by a marked initial cognitive offloading.
- Phase 3: Critical Mediation (Human-in-the-loop). Introduction of the artifact (the rubric). Students redesigned their proposals by applying cognitive oversight criteria, reclaiming evaluative judgment over the automated design. This phase served as an iterative refinement cycle where student feedback on rubric usability was used to calibrate the final evaluative criteria.
- Phase 4: Evaluation. Administration of the post-test and collection of final artifacts for technical audit.

These iterative phases are consistent with the design-based research cycles adopted in the broader international research initiative Horizon Europe STEAMbrace, where educational interventions are progressively refined through cycles of implementation, evaluation, and redesign in authentic learning contexts.

3.5 Data analysis

A mixed-methods approach was employed to capture the complexity of the phenomenon. This triangulation allows for a comprehensive understanding of how the analytical mediation model facilitates the transition from mere technical interaction to the exercise of professional epistemic agency.

- Quantitative Analysis: Hypothesis testing was conducted using paired-samples *t*-tests or the Wilcoxon signed-rank test, depending on the normality of the distribution (assessed via Shapiro–Wilk). Effect size was calculated using Cohen's d to determine the magnitude of the pedagogical impact. In addition, independent samples *t*-tests or ANOVA were used to compare performance between academic levels (first-year vs. second-year) to identify potential performance ceilings related to Pedagogical Content Knowledge (PCK).
- Qualitative Analysis: To ensure a systematic and unbiased evaluation of the narrative evidence, a thematic content analysis was conducted on a stratified random subsample of $n = 40$ student portfolios selected from the full dataset ($N = 195$). Stratification was based on the quantitative performance quartiles of the final design quality scores, ensuring an equal representation of 10 portfolios from each quartile (high, mid-high, mid-low, and low performance). The qualitative material subjected to analysis integrated both the technical attributes of the AI-generated/rubric-refined learning situations and the students' corresponding text reflections. A hybrid deductive-inductive coding approach was deployed. Deductive codes were anchored in the core theoretical constructs (e.g., *epistemic agency, cognitive friction, PCK constraints*), while inductive sub-codes emerged dynamically from the text corpus (e.g., *supervision fatigue, structural compliance*). To guarantee psychometric rigor, two researchers coded the material independently. The initial inter-rater reliability analysis yielded a high level of agreement (Cohen's $\kappa = 0.82$); any remaining coding discrepancies or boundary disagreements were subsequently resolved through consensual deliberation sessions until 100% agreement was achieved."
- Mixed-Methods Integration: Following an explanatory sequential mixed-methods paradigm, the qualitative findings were systematically integrated with the quantitative results. The narrative themes identified in Section 4.4 were directly utilized to contextualize and explain the underlying cognitive and metacognitive mechanisms driving the highly significant statistical variances (*F*-values and *t*-tests) observed across the design phases.

4 Results

This section presents the study's findings, organized according to the research questions. The analysis combines quantitative evidence derived from the evaluation of didactic proposals and the pre-test and post-test questionnaires administered to the students. Results are structured into three key areas: (1) the evolution of pedagogical quality in the didactic proposals throughout the intervention process, (2) the effect of rubric-based mediation on the improvement of pedagogical designs, and (3) changes in student perceptions regarding the educational use of generative artificial intelligence.

4.1 Pedagogical quality and design descriptives (RQ1)

The evaluation of the internal consistency of the mediation rubric yielded Cronbach's α coefficients of .86 in the pre-test and .89 in the post-test ($N = 195$), confirming the instrument's robustness and stability for measuring pedagogical design quality.

Prior to the contrast analysis, the evolution of performance across the three intervention phases is presented (Table 1). This progression illustrates the transition from a baseline human-led design (Phase 1), through the exploration of GenAI affordances with a tendency toward cognitive offloading (Phase 2), to the analytical mediation phase (Phase 3), aimed at the reclaiming of evaluative judgment by the pre-service teacher.

Data analysis reveals that the Competency-based Objectives dimension shows the greatest absolute improvement (+2.50 points) and the lowest final variance ($SD = 0.41$). This convergence of results confirms that GenAI possesses high efficacy in syntactic structuring tasks, which tend to stabilize rapidly following the application of the scaffolding provided by the rubric. These data suggest the existence of a ceiling effect in purely organizational tasks, where the tool reaches its maximum operational potential under minimal teacher supervision.

In contrast to structural dimensions, Interdisciplinary/STEAM Integration maintains a high standard deviation ($SD = 0.92$) in the final phase of the intervention. The fact that 17.4% of subjects failed to reach the optimal level highlights a critical transfer gap. The results indicate that AI affordances encounter a performance ceiling when faced with high-cognitive-demand tasks requiring complex epistemological synthesis. This phenomenon suggests that analytical scaffolding is insufficient without a solid foundation of prior Pedagogical Content

Knowledge (PCK), making interdisciplinary complexity a barrier to critical oversight of the algorithm.

Significantly, the transition from Phase 2 (autonomous AI use) to Phase 3 (rubric-mediated use) reduced result dispersion by an average of 35%. This contraction of variance should not be interpreted merely as a grade improvement, but as the consolidation of the rubric as a cognitive oversight artifact. By mitigating algorithmic drift and standardizing pedagogical quality criteria, the intervention enables students to reclaim and exercise their epistemic agency during the instructional design process.

Prior to conducting the inferential analysis, normality assumptions were tested using the Shapiro–Wilk test ($p > .05$), and homogeneity of variance was verified across all dimensions. For the repeated measures ANOVA, Mauchly's test of sphericity was applied. In cases where a violation of this assumption was detected (Mauchly's W, Greenhouse-Geisser $\epsilon < .75$), Greenhouse-Geisser corrections were employed. All *post-hoc* pairwise comparisons between the three design phases were subjected to strict Bonferroni-adjusted significance corrections to mitigate the risk of Type I error inflation (p_{adj}).

This significant quantitative progression across all five dimensions of pedagogical quality reflects more than a mere technical optimization; it denotes a fundamental cognitive shift. As further detailed in the qualitative triangulation (see Section 4.4, Theme 2), the statistical leap from Phase 2 to Phase 3 directly corresponds to the students' transition from passive acceptance of automated outputs to active, rubric-mediated interrogation of the AI. The narrative corpus corroborates that the highly significant F-values observed in dimensions like pedagogical coherence ($F = 342.15$, $p < 0.001$) are anchored in the students' conscious application of the rubric as a deliberative brake, converting automated text into highly contextualized curricular designs.

4.2 Inferential analysis of the mediation effect

For the inferential analysis, assumptions of normality and homogeneity were tested. A repeated measures ANOVA was conducted. Sphericity was verified using Mauchly's test; in dimensions where the assumption was violated ($p < 0.05$), the Greenhouse-Geisser correction was applied.

The results (Table 2) confirm a significant main effect of the Intervention Phase factor across all dimensions ($p < 0.001$), with effect sizes η_p^2 denoting a high-magnitude pedagogical impact.

Pairwise comparison analysis using Bonferroni adjustment reveals critical findings for the validation of the intervention model. Across all dimensions analyzed, the transition from raw GenAI use (Phase 2) to rubric-mediated use (Phase 3) yielded significant increases ($p < 0.01$). These results suggest that autonomous AI usage reaches a technical performance ceiling that is only surpassed when students activate their pedagogical judgment through the mediation artifact, transforming their interaction with the machine from passive cognitive offloading into an exercise in reflective design.

In terms of statistical power, the Competency-based Objectives dimension recorded the highest F-statistic in the study ($F = 401.82$). This data indicates that the cognitive oversight rubric is particularly effective at mitigating the intrinsic tendency of LLMs to produce generic, decontextualized outputs. Analytical mediation compels pre-service teachers to undertake a technical re-specification of learning goals, exercising an epistemic agency that reclaims control over the pedagogical intentionality of the design against algorithmic standardization.

Conversely, the Interdisciplinary/STEAM Integration dimension presented the most moderate magnitude of change within the model ($\eta = 0.33$). *Post-hoc* analysis indicates that the difference between Phase 2 and Phase 3, while statistically significant, is smaller compared to the structural dimensions of the curriculum. This phenomenon reflects the high cognitive demand required for the synthesis of knowledge areas, where the scaffolding provided by the rubric encounters greater transfer barriers if the students lack a consolidated prior conceptual foundation. This finding reinforces the necessity of prioritizing disciplinary foundations as a prerequisite for the critical oversight of AI.

4.3 Reconfiguring teacher identity: evaluative judgement and the self-efficacy paradox (RQ2)

To evaluate the intervention's impact on the subjective dimensions of pre-service teachers, an inferential analysis was conducted using paired-samples t-tests. This analysis seeks to determine whether analytical mediation—conceptualized as external scaffolding—alters students' perception of competence and their ethical disposition. Prior to this, the normality assumptions for the differences were confirmed using the Shapiro–Wilk test ($p > 0.05$). The results (Table 3) reveal a significant shift across all variables, highlighting a transition from technical operability toward the strengthening of professional judgement.

Calculation of Cohen's d to measure the magnitude of change yielded values exceeding 0.80 in the Pedagogical Judgment and Practical Knowledge dimensions, constituting a large effect size by disciplinary standards. Notably, the $t = 35.12$ value for Pedagogical Judgment represents the greatest magnitude of change observed in the study. This massive shift indicates that rubric-based mediation functioned as a catalyst for evaluative judgment. Specifically, it enabled students to transition from a heuristic confidence in algorithmic output toward expert validation. This process represents more than a mere technical optimization; it constitutes a genuine reclamation of pedagogical agency. Ultimately, pre-service teachers abandon the role of passive consumers to position themselves as critical auditors of automated processes.

TABLE 3 Paired-samples t-test results for perceived self-efficacy and professional judgment ($N = 195$).

Dimension	Pre-test (M/SD)	Post-test (M/SD)	t	p
Academic AI usage	2.15/0.82	4.38/0.54	28.45	< 0.001
AI self-efficacy	2.84/1.05	3.92/0.91	9.78	< 0.001
Practical knowledge	2.32/0.88	4.15/0.62	21.34	< 0.001
Pedagogical judgment	1.98/0.76	4.56/0.48	35.12	< 0.001
Ethical awareness	3.05/1.12	4.41/0.74	12.67	< 0.001

It is highly significant that Self-efficacy exhibited the most moderate growth in the model ($t = 9.78$) alongside the highest final dispersion ($SD = 0.91$). From a Self-Regulated Learning (SRL) perspective, this suggests that critical training activated more demanding metacognitive monitoring processes (Panadero, 2017; Pintrich, 2000). This empirical manifestation, which we term the "self-efficacy paradox," demonstrates that unmediated interaction with generative AI tools creates an initial "illusion of competence" (Koriat, 1997), where superficial output fluency is mistaken for actual task mastery. Students evolved from unfounded security toward a phase of conscious competence, where their perception of their own ability is recalibrated upon recognizing the ethical and technical complexity of the task. This critical realism serves as an indicator of professional maturity; the subject understands that AI oversight requires a deliberative effort previously ignored (Zimmerman and Moylan, 2009). As Bandura (1997) noted, miscalibrated self-efficacy occurs precisely when authentic, objective feedback loops are absent. Consequently, this phenomenon provides direct empirical evidence of the dismantling of the Dunning-Kruger effect (Kruger and Dunning, 1999), where 16.5% of subjects experienced a necessary competence crisis to mitigate algorithmic complacency. This recalibration aligns with the transition from "unconscious incompetence" to "conscious competence" (or even "conscious incompetence" regarding AI limitations), a vital step for professional expertise where metacognitive accuracy is prioritized over speed.

Finally, the recorded increase in Ethical Awareness ($M = 4.41$) indicates a paradigm shift in human-machine interaction. Students abandon the view of AI as a substitute oracle, positioning it instead as an interlocutor under constant oversight. In this context, ethics ceases to be a peripheral theoretical concept and is transformed into an operational competence: the ability to detect the erosion of authorship and protect the integrity of educational design against the inertia of automation. This consolidation of ethical judgment ensures that technology is integrated as a supporting resource rather than a replacement for professional pedagogical judgment.

The statistical fluctuations in perceived self-efficacy capture the internal tension characterizing the development of early professional teacher identity in the algorithmic era. The quantitative contraction observed after autonomous AI usage finds deep empirical explanation in the participants' personal reflections (see Section 4.4). The narrative data reveals that the drop in self-efficacy is not an indicator of pedagogical regression, but rather a healthy "shattering" of naive confidence. The empirical qualitative evidence confirms that when faced with objective criteria, pre-service teachers renegotiate their sense of agency, shifting from technical fluency to an acute awareness of the professional responsibilities inherent to instructional design.

4.4 The semantic transition process: qualitative triangulation (RQ3)

To understand the cognitive mechanism behind the quantitative shifts in self-efficacy and design quality, a thematic analysis of participants' open-ended reflections was conducted. The semantic transition from uncritical reliance on AI to reflective pedagogical agency is structured around three prominent themes.

Initial reflections revealed that autonomous AI usage (Phase 2) generated an inflated sense of competence due to the immediate, grammatically perfect output of the LLM. However, this technical fluency masked a lack of pedagogical depth. The qualitative data captures the exact moment this illusion shatters, explaining the statistical drop in self-efficacy. As **Subject 84** noted: "*At first, I thought the AI lesson plan was perfect because it sounded highly professional and was generated in seconds. But when I had to evaluate it against the rubric, I realized it lacked real pedagogical coherence; the objectives were detached from the actual steps. It made me feel less confident initially, but more aware of what a real design requires.*" This aligns with the "Dunning-Kruger" effect highlighted by Kruger and Dunning (1999) and the warnings of Selwyn (2024) regarding the seductive nature of automated text, where superficial quality conceals pedagogical emptiness.

The introduction of the analytical rubric in Phase 3 introduced what participants described as productive "cognitive friction." The rubric acted as a human-in-the-loop mechanism that forced students to slow down and shift from technical operation to evaluative judgment. **Subject 159** reflected: "*The rubric forced me to interrogate the AI. I couldn't just accept its first response. I had to create a prompt loop, correcting its structure until the competency framework aligned with the Basque curriculum. The rubric was the guide that allowed me to take control back from the machine.*" This narrative evidence supports the quantitative surge in design quality, illustrating how analytical scaffolds act as metacognitive tools (Pintrich, 2000; Zimmerman, 2000) that prevent cognitive offloading and foster epistemic agency (Bearman and Ajjawi, 2023).

Despite the support provided by both the AI and the rubric, complex interdisciplinary integration (such as STEAM) presented a performance ceiling. The qualitative data suggests that when students lacked a solid foundation of Pedagogical Content Knowledge (PCK), the technology could not substitute for content mastery, leading to automated fatigue. **Subject 31** stated: "*In the general sections, the AI and the rubric helped a lot. But when it came to integrating science and art in the STEAM section, I felt stuck. The AI kept giving generic ideas, and since I don't master the STEAM methodology myself, I didn't know how to instruct the AI to fix it. It was exhausting to supervise an AI when you are not sure of the content yourself.*" This experience explains why STEAM integration achieved lower quantitative scores compared to other dimensions. It confirms Shneiderman's (2022) Human-Centered AI paradigm: AI amplification is only effective when anchored in robust human expertise; otherwise, it induces supervision fatigue and cognitive debt (Sweller, 1988; Tao et al., 2026).

5 Discussion

5.1 Evaluative judgment vs. cognitive offloading

Quantitative results suggest that the rubric-mediated intervention for cognitive oversight is associated with a substantial improvement in pedagogical design quality. The recording of a large effect size in critical dimensions such as

Pedagogical Judgment and Practical Knowledge ($d > 0.80$) is not an isolated finding; rather, it aligns with the most recent meta-analytic evidence. In this regard, global meta-analyses on the use of LLMs in higher education report that GenAI applications, when properly integrated, achieve impact magnitudes exceeding the average for traditional technological interventions (Mo et al., 2026; Hedges' $g = 1.14$).

This convergence of data suggests that the superiority of Phase 3 (AI + rubric) over Phase 2 (AI alone) does not merely reflect an increase in students' technical proficiency, but rather the activation of higher-order monitoring processes. When contrasting our results with international standards for AI in education, it is validated that the proposed mediation model acts as a multiplier of AI affordances, transforming the raw potential of the algorithm into a pedagogically sound professional performance. This finding reinforces the thesis that teaching proficiency in the digital era does not stem from the mere accumulation of information, but from the synergy between competency development and the capacity to act flexibly and critically with knowledge (Bilbao et al., 2021). Through this lens, the rubric acts as the engine that triggers "deep learning"—for the human, ironically, not just the machine—enabling pre-service teachers to move beyond passively receiving AI-generated data to interpret, contextualize, and relate it to their professional environment autonomously. This development is likely linked not to the standalone capability of the language model itself, but to the structured rigor of the scaffolding that prompts the student to exercise constant analytical oversight over the generated output. Active monitoring is, ultimately, the safeguard mechanism that prevents cognitive offloading, ensuring that the intellectual effort of didactic transposition remains the responsibility of the pre-service teacher and is not passively delegated to the AI system. As Castañeda and Williamson (2021) argue, educational technology must be understood as a complex assemblage where critical research is needed to reclaim human agency. The rubric, in this study, serves as the methodological toolbox required to navigate this automated landscape without losing the teacher's sovereign control.

This need for analytical mediation aligns with the findings of Hauk and Soujon (2026), who demonstrate that the reliability of LLMs in evaluating and generating lesson plans is inconsistent without a human expert standard to guide the interaction. As in their study, our data suggest that AI does not replace teacher judgment; instead, it requires an external evaluation structure—in our case, the rubric—to achieve professionally acceptable levels of quality. As Shi et al. (2026) indicate in their recent systematic review, the impact of LLMs on academic performance is maximized when they are utilized as scaffolding tools that actively promote students' cognitive capabilities.

However, the exceptional magnitude of these empirical gains—particularly the statistical shift observed in pedagogical judgment—demands a highly cautious and nuanced interpretation. These prominent effect sizes should not be interpreted as an unconfounded, absolute transformation in students' deep cognitive reasoning. Instead, they must be critically understood as a reflection of several intersecting contextual factors inherent to the intervention ecosystem. First, a profound assessment alignment existed between the analytical scaffold utilized during training and the final outcome measures. Because the rubric functioned simultaneously as a learning guide and the evaluation metric, a portion of this statistical progression inevitably reflects students strategically learning "how to satisfy the rubric" through structural compliance. Second, the intensive, ongoing instructor guidance provided across the DBR cycles, combined with the participants' rapid familiarity with the explicit descriptors, acted as a powerful performance accelerator. Finally, the potential for subtle rater expectancy effects during the Technical Audit phase cannot be entirely ruled out. Acknowledging these dynamics does not invalidate the model's utility, but it recontextualizes the large t-values as successful indicators of instructional alignment and scaffold execution rather than isolated, unmediated cognitive growth.

5.2 Resistance in STEAM complexity and the AI performance ceiling

The persistent high variability in the STEAM dimension ($SD = 0.92$) constitutes one of the most significant findings of this study. While GenAI exhibits exceptional affordances for syntactic structuring and the formal organization of the curriculum, its synthesis capacity encounters a performance ceiling when faced with high-cognitive-demand tasks that require authentic interdisciplinary integration. This limitation is consistent with evidence from Hauk and Soujon (2026), who demonstrate that while language models can emulate the structure of a lesson plan, their reliability drops drastically when evaluating or generating content that demands deep, contextualized pedagogical reasoning. This resistance highlights that the algorithm can assist in the formal architecture of design but remains incapable of substituting the teacher's Pedagogical Content Knowledge (PCK). This limitation echoes Sahlberg's (2024) warning that the "AI revolution" in schools risks automating mediocrity if it is not guided by high-level professional capital. Our results confirm that without a solid human-led pedagogical foundation, the machine merely produces formally correct but semantically shallow structures.

The identified gap suggests that the scaffolding provided by the rubric encounters a critical limit in the subject's prior competence. If the teacher lacks solid conceptual frameworks regarding the STEAM approach, technological mediation becomes insufficient to guarantee expert oversight. This result reinforces the thesis that teacher education must transcend mere prompt engineering to prioritize the strengthening of epistemological and disciplinary foundations (Yoon et al., 2026). From this perspective, using AI in contexts of low disciplinary competence can lead to what Tao et al. (2026) describe as cognitive debt: a latent risk where the user uncritically accepts sub-optimal AI solutions, thereby mortgaging their own long-term professional development.

Without this foundation, the teacher risks becoming a blind supervisor: a user capable of identifying superficial anomalies in the output but unable to perform the critical synthesis and professional judgment that the complexity of interdisciplinary education demands. This dichotomy is reflected in the usage profiles identified by Hugerth and Hugerth (2026), who distinguish between efficiency-oriented and knowledge-oriented AI use. Our findings suggest that, in the absence of a robust PCK, the mediation model may be used merely as an efficiency shortcut, failing in its purpose to promote true epistemic agency in the design of STEAM learning situations. This tension reflects the "limits of AI" identified by Selwyn (2024), where the

uncritical delegation of educational tasks to algorithms threatens to deprofessionalize teaching. The mediation model acts here as a barrier against this erosion of authority.

5.3 The self-efficacy paradox and the cost of agency

The qualitative shift in perceived self-efficacy—which significantly increased quantitatively but evolved into a more cautious, reflective, and critically calibrated state—contrasted with the notable increase in actual performance, constitutes the central paradox of this research. This phenomenon aligns with the user profile taxonomy proposed by Hugerth and Hugerth (2026), who distinguish between efficiency-seekers and knowledge-seekers. The resistance detected in students suggests that the rubric's analytical scaffolding imposes a deliberative load that clashes with the predisposition toward algorithmic shortcuts. By forcing a transition from the pursuit of immediate results toward epistemic deepening, the mediation model generates a necessary friction that dismantles the beginner's naive confidence, compelling a vital departure from the Dunning-Kruger effect in favor of grounded professional judgment.

On the other hand, the testimony of Subject 31—stating that "it's a chore to use the rubric because it takes longer than doing it myself"—provides empirical evidence of the transactional cost involved in maintaining human agency. Through the lens of Hugerth and Hugerth (2026), this supervision fatigue is characteristic of efficiency-oriented profiles, for whom AI functions as a cognitive offloading mechanism that the rubric effectively neutralizes. This "laziness" is, in fact, a resistance to abandoning metacognitive passivity (Tao et al., 2026), revealing that contemporary teacher education must focus not only on technical skills but also on ethical resilience against the seduction of algorithmic immediacy.

Safeguarding this effort is essential for achieving SDG 4 (Quality Education). If teacher education programs succumb to the "seduction of immediacy," they risk producing graduates who lack the epistemic integrity needed to ensure inclusive and rigorous learning in an algorithmic era. Ultimately, this critical recalibration of self-efficacy, where quantitative gains are accompanied by a rigorous and realistic self-assessment, should be interpreted as an indicator of professional maturity and a success of the critical mediation model. Students have moved beyond unconscious incompetence to confront the actual complexity of professional teaching judgment. Teaching how to resist algorithmic speed is, paradoxically, one of the most urgent learning objectives in the GenAI era. Only through this critical realism is it possible to transform the interaction with technology from a passive delegation into a reflective scaffolding that strengthens epistemic agency and teaching identity instead of eroding them.

6 Implications

6.1 From instrumental literacy to algorithmic sovereignty: theoretical and policy implications

The findings of this research suggest that the integration of GenAI in Higher Education must transcend the acquisition of specific technical skills. Data demonstrate that design quality does not emanate from the intrinsic affordances of the algorithm, but rather from mediation structures that make pedagogical quality criteria explicit. From a theoretical perspective, this study expands the HCAI (Human-Centered Artificial Intelligence) framework by demonstrating that Human-in-the-loop design is not merely a technical configuration, but an ethical and cognitive disposition that requires explicit training.

In terms of educational policy, digital competence frameworks for teachers, such as DigCompEdu, must urgently evolve. It is no longer sufficient to evaluate the ability to create content with AI; it is imperative to institutionalize pedagogical auditing capabilities over automated systems. Initial teacher education must prioritize this evaluative judgment as a defense mechanism against what Tao et al. (2026) define as cognitive debt: the risk of future teachers losing their critical thinking schemas by delegating didactic synthesis to the algorithm. The rubric, in this context, acts as a small-scale algorithmic governance artifact that returns intellectual sovereignty to the classroom.

6.2 Orchestrating human-in-the-loop designs and iterative cycles

To mitigate the risk of cognitive offloading, instructional design must evolve toward socio-cognitive models where AI acts as a prototype generator and the teacher assumes the role of critical evaluator. This Human-in-the-loop approach requires that initial training activities follow an iterative, rather than linear, logic. Under this paradigm, AI is used to produce initial drafts that must undergo mandatory cycles of revision and comparison through analytical scaffolding, ensuring that human agency remains at the center of pedagogical decision-making.

6.3 The rubric as an oversight and evaluative judgement artifact

The transition from uncritical technological fascination to expert oversight requires formative assessment tools to be reconfigured as cognitive oversight artifacts. The use of analytical rubrics in this study has proved to be a fundamental catalyst for the development of evaluative judgement. By imposing explicit constraints and quality criteria, the rubric compels students to interrogate the algorithm based on sound didactic principles, preventing the superficial eloquence of language models from replacing the necessary reflection on design and educational practice.

6.4 Managing supervision fatigue and the cost of agency

Finally, it is imperative to consider the cognitive cost involved in human oversight within automated environments. As evidenced by the resistance detected in cases such as Subject 31, critical mediation generates a deliberative friction

that may be perceived as a barrier to temporal efficiency. Instructional designers must, therefore, balance scaffolding demands with student autonomy to prevent supervision fatigue. Teacher education must acknowledge that maintaining human agency is an inherently slower and more costly process, yet one that is essential for preserving quality and ethics in the algorithmic era.

7 Conclusions and limitations

7.1 Superiority of mediation and the syntax–semantics dichotomy

This research indicates that interaction paired with analytical rubrics is associated with higher design quality outcomes compared to the unmediated autonomous use of GenAI within the context of initial teacher education. The results reveal a critical dichotomy in the algorithm's capabilities: while AI demonstrates high efficacy in educational syntax—defined as the formal and organizational structuring of design—its ability to generate situated and coherent semantics depends strictly on the user's pedagogical oversight. In this sense, mediation not only improves the final product but ensures that the design responds to actual pedagogical needs rather than mere automated content generation.

7.2 The self-efficacy paradox as a metric of training success

A fundamental finding of this study is the validation of the self-efficacy paradox as a quality indicator in technological training. Contrary to the traditional view that associates learning with a simple linear increase in confidence, this work suggests that effective critical training is that which recalibrates students' initial naive confidence. By transforming an illusory perception of competence into a cautious, reflective, and critically calibrated self-efficacy, the proposed mediation model prepares pre-service teachers to more rigorously confront the biases and limitations intrinsic to algorithmic systems. This recalibration is essential for fostering the human flourishing and epistemic integrity required by SDG 4, ensuring that future teachers act as sovereign agents in democratic educational settings.

7.3 Study limitations: scope and ecological validity

Despite the robustness of the DBR (Design-Based Research) design, it is necessary to acknowledge several limitations that define the scope and interpretation of these results. First, the use of non-probabilistic convenience sampling in a single university context suggests that the generalization of findings should be approached with caution. Second, this study faces an ecological validity challenge, as the analysis focused exclusively on the didactic planning phase; evaluating the impact of these proposals in actual classroom practice remains a pending task. Third, a primary methodological limitation is the lack of a true experimental or quasi-experimental control group, which prevents the research design from completely isolating historical or maturation effects from the net impact of the intervention. Furthermore, as previously discussed, the massive statistical effect sizes may be partially inflated by the structural alignment of the tool, potential rater bias, or students adapting their outputs simply to satisfy the rubric's explicit demands. Finally, the lower efficacy of mediation detected in the STEAM dimensions confirms that technological scaffolding does not compensate for a lack of prior Pedagogical Content Knowledge (PCK), reinforcing the need for high-level professional capital as a prerequisite for AI oversight.

7.4 Outlook and future research directions

As avenues for continuity, longitudinal studies are proposed to analyze whether the critical realism developed by students persists after their entry into the professional world. Furthermore, it is of great interest to explore the impact of AI-mediated peer review. This line of research would help determine whether socio-technical collaboration is capable of mitigating the cognitive fatigue and transactional costs identified in this study, thereby optimizing teacher oversight processes in environments with high algorithmic demand.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin because the data were collected as part of the routine educational curriculum and analyzed retrospectively in an anonymous format, in accordance with local institutional guidelines for educational research.

Author contributions

NB-Q: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. ET: Formal analysis, Investigation, Methodology, Writing – review & editing. AR: Formal analysis, Investigation, Methodology, Writing – review & editing. AL: Formal analysis, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This study was conducted within the framework of the Horizon Europe STEAMbrace project, funded by the European Commission under Grant Agreement No. 101094874. The research contributes to the empirical examination of pedagogical models that integrate generative artificial intelligence (GenAI) as a mediating tool for teacher education, focusing on the design and validation of innovative STEAM educational practices supported by digital technologies.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. New York: W. H. Freeman.

Bearman, M., and Ajjawi, R. (2023). Learning to work with the black box: pedagogy for a world with generative AI. *Br. J. Educ. Technol* 54, 437–453. doi: 10.1111/bjet.13339

Bilbao, N., Garay, U., Romero, A., and López de la Serna, A. (2021). The European competency and the teaching for understanding frameworks: creating synergies in the context of initial teacher training in higher education. *Sustainability* 13, 1810. doi: 10.3390/su13041810

Caspari-Sadeghi, S. (2026). AI Literacy for teacher educators: a holistic curriculum for capacity-building in higher education. *Front. Educ* 11, 1745768. doi: 10.3389/feduc.2026.1745768

Castañeda, L., and Williamson, B. (2021). Assembling new toolboxes of methods and theories for innovative critical research on educational technology. *J. New Approaches Educ. Res* 10, 1–14. doi: 10.7821/naer.2021.1.703

Chaudhry, M. A., Saleem, H., and Javaid, M. (2023). ChatGPT: a review of opportunities and challenges for education. *J. Artif. Intell. Appl* 2, 1–15. doi: 10.1016/j.aiia.2023.100021

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). Opinion paper: "so what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manage* 71, 102642. doi: 10.1016/j.ijinfomgt.2023.102642

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Hauk, D., and Soujon, N. (2026). How reliable are large language models in analyzing the quality of written lesson plans? A mixed-methods study from a teacher internship program. *Comput. Educ. Artif. Intell* 10, 100538. doi: 10.1016/j.caeai.2026.100538

Hugerth, M. W., and Hugerth, L. W. (2026). Seeking knowledge or efficiency: profiling students' AI-use through survey-based latent class analysis. *Comput. Educ. Artif. Intell* 10, 100531. doi: 10.1016/j.caeai.2026.100531

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ* 103, 102274. doi: 10.1016/j.lindif.2023.102274

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *J. Exp. Psychol. Gen* 126, 60–70. doi: 10.1037/0096-3445.126.1.60

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Li, Z., and Li, H. (2026). Unveiling the pedagogical potential of large language models for critical and creative thinking: a systematic review. *Front. Educ* 11, 1816306. doi: 10.3389/feduc.2026.1816306

Luckin, R. (2024). *AI for Learning: How to Help Your Students Learn with and About AI*. London: Sage.

McKenney, S., and Reeves, T. C. (2018). *Conducting Educational Design Research*. 2nd ed. New York: Routledge. doi: 10.4324/9781315105642

Mishra, P., and Koehler, M. J. (2006). Technological pedagogical content knowledge: a framework for teacher knowledge. *Teach. Coll. Rec* 108, 1017–1054. doi: 10.1111/j.1467-9620.2006.00684.x

Mo, F., Huang, J., Yang, Y., Özen, Z., Maeda, Y., and Olenchak, F. R. (2026). Undergraduate students' learning outcomes with ChatGPT: a meta-analytic study. *Comput. Educ. Artif. Intell* 10, 100536. doi: 10.1016/j.caeai.2026.100536

Molenaar, I. (2022). Towards hybrid-human-AI learning technologies. *Eur. J. Educ* 57, 632–645. doi: 10.1111/ejed.12527

Panadero, E. (2017). A review of self-regulated learning: six models and four directions for research. *Front. Psychol* 8, 422. doi: 10.3389/fpsyg.2017.00422

Pintrich, P. R. (2000). "The role of goal orientation in self-regulated learning," in *in Handbook of Self-regulation*, eds. M. Boekaerts, P. R. Pintrich, and M. Zeidner (San Diego: Academic Press), 451–502. doi: 10.1016/B978-012109890-2/50046-9

Rind, I. A. (2026). Conceptualizing the impact of AI on teacher knowledge and expertise: a cognitive load perspective. *Educ. Sci* 16 (1), 57. doi: 10.3390/educsci16010057

Rind, I. A., Bhatti, J., and Chellappan, K. (2026). From blind trust to critical inquiry: epistemic beliefs and student engagement with ChatGPT in higher education. *J. Kejuruteraan* 38 (1), 1–14. doi: 10.17576/jkukm-2026-38(1)-01

Rind, I. A., and Qureshi, M. A. (2026). AI-Mediated Teaching in K-12 Classrooms. Preprints. doi: 10.20944/preprints202604.0660.v1

Sahlberg, P. (2024). Has the AI revolution arrived in schools? *J. Prof. Cap. Commun* 9, 1–10. doi: 10.1108/JPCC-12-2023-0089

Selwyn, N. (2024). On the limits of artificial intelligence (AI) in education. *Nord. Tidsskr. Pedag. Kritikk* 10, doi: 10.23865/ntpk.v10.6062

Shi, Y., Yu, K., Dong, Y., and Chen, F. (2026). Large language models in education: a systematic review of empirical applications, benefits, and challenges. *Comput. Educ. Artif. Intell* 10, 100529. doi: 10.1016/j.caeai.2026.100529

Shneiderman, B. (2022). *Human-centered AI*. Oxford: Oxford University Press. doi: 10.1093/oso/9780192845290.001.0001

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognit. Sci* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Tao, S., Lan, M., Wang, M., and Li, H. (2026). Potential risks of generative artificial intelligence integration into K-12 education: a scoping review. *Comput. Educ. Artif. Intell* 10, 100561. doi: 10.1016/j.caeai.2026.100561

UNESCO. (2023). *Guidance for Generative AI in Education and Research*. Paris: UNESCO Publishing. doi: 10.54675/ASBC9150

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press.

Yañez-Perea, A., Bilbao-Quintana, N., and López de la Serna, A. (2025b). "Inteligencia artificial para el desarrollo metacognitivo del alumnado en espacios

híbridos: aportes al diseño del espacio 'pentsatu' de IkasLab," *Horizontes Educativos con Inteligencia Artificial*, ed. Dykinson(Madrid: Dykinson), 32–44.

Yañez-Perea, A., Bilbao-Quintana, N., and López-De la Serna, A. (2025a). Delphi validation of a rubric for IkasLab spaces for active and global learning. *Educ. Sci* 15, 1610. doi: 10.3390/educsci15121610

Yoon, I., Ryu, J., Kim, K., and Kim, I. (2026). What AI–digital competencies should teachers develop throughout their careers? Designing a career-responsive framework through a Delphi study. *Front. Educ* 11, 1745059. doi: 10.3389/feduc.2026.1745059

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int. J. Educ. Technol. High. Educ* 16, 39. doi: 10.1186/s41239-019-0171-0

Zimmerman, B. J. (2000). "Attaining self-regulation: a social cognitive perspective," in *in Handbook of Self-regulation*, eds. M. Boekaerts, P. R. Pintrich, and M. Zeidner (San Diego: Academic Press), 13–39. doi: 10.1016/B978-012109890-2/50031-7

Zimmerman, B. J., and Moylan, A. R. (2009). "Self-regulation: where metacognition and motivation intersect," in *Handbook of Metacognition in Education*, eds. D. J. Hacker, J. Dunlosky, and A. C. Graesser (New York: Routledge), 299–315.