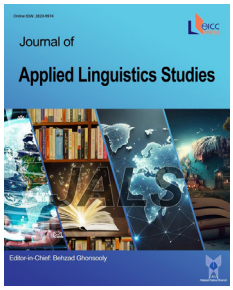

Volume 5, Issue 2, 052503 (174-187)

Journal of Applied Linguistics Studies (JALS)

https://doi.org/10.57647/jals.2025.0502.03

AI-Assisted Lesson Planning and Teacher Development: A Mixed-Methods Study on Critical Thinking and Creativity

Roghaye Torki[1], Fariba Rahimi Esfahani[1]*, Farshad Kiyoumarsi[2]

[1]*Department of English, ShK.C., Islamic Azad University, ShahreKord, Iran*
[2]*Department of Computer, ShK.C., Islamic Azad University, ShahreKord, Iran*

***Corresponding author:** fariba.rahimi@iau.ac.ir

Original Article

Received:
2025-08-31

Revised:
2025-10-04

Accepted:
2025-11-11

Published in Issue:
2025-12-30

Abstract:

The rapid integration of generative Artificial Intelligence, such as ChatGPT, in education is reshaping teachers' professional practices. However, its influence on teachers' higher-order cognitive processes during instructional design remains insufficiently understood. This study examined the impact of ChatGPT-assisted lesson planning on EFL teachers' critical thinking, creativity, and overall instructional design quality. Using an explanatory sequential mixed-methods quasi-experimental design, the research combined quantitative and qualitative strands to determine how AI-supported planning influenced teachers' cognitive engagement. Forty-two in-service EFL teachers were purposively selected and randomly assigned to an experimental group, that received structured training on integrating ChatGPT into lesson planning, and a control group, that planned lessons using conventional methods. Participants completed pretest and posttest lesson-planning tasks evaluated through three validated instruments: The Critical Thinking Rubric, the Creativity Rubric, and the Lesson Plan Quality Scale. Repeated-measures MANOVA results revealed significant time × group interactions across all three domains, with the experimental group demonstrating substantially higher posttest gains. To enrich these findings, semi-structured interviews with a subsample of twelve teachers explored their planning experiences and cognitive processes. Thematic analysis showed that AI-supported planning enhanced teachers' ability to justify instructional decisions, expanded their creative repertoires, and fostered more deliberate pedagogical reasoning, while also highlighting the importance of critical engagement to avoid passive reliance on AI-generated content. Collectively, the results suggest that when used interactively and reflectively, ChatGPT can function as a productive cognitive scaffold supporting teachers' instructional design. Implications include the need for targeted professional development to ensure that AI is integrated responsibly and in ways that strengthen rather than diminish teachers' pedagogical agency.

Keywords: Artificial Intelligence; ChatGPT; Creativity; Critical Thinking; Lesson Planning

Cite this article: Torki, R., Rahimi Esfahani, F., Kiyoumarsi, F. (2025) AI-Assisted Lesson Planning and Teacher Development: A Mixed-Methods Study on Critical Thinking and Creativity. Journal of Applied Linguistics Studies, 5(2), 174-187. https://doi.org/10.57647/jals.2025.0502.03

INTRODUCTION

Advances in artificial intelligence (AI) have introduced new ways to support teachers' instructional planning and reshape the cognitive demands of designing effective lessons, with tools such as ChatGPT gaining prominence for generating initial ideas, reorganizing content, and proposing alternative activity sequences that may streamline preparation and expand pedagogical options. Early evidence suggests that these tools help teachers explore multiple instructional pathways more efficiently and approach complex planning with greater flexibility

(Karakas & Yeşilyurt, 2023; Li & Ni, 2024), yet they also revive concerns about whether AI truly enhances teachers' cognitive engagement or simply simplifies surface-level performance in tasks that traditionally rely on analytical and creative reasoning. Empirical findings reflect this tension. Research shows that structured interaction with ChatGPT can broaden teachers' exposure to diverse instructional models and support more flexible creative ideation (Zhai et al., 2024; Zhua et al., 2024), while other studies caution that dependence on AI-generated suggestions may reduce opportunities for critical evaluation, contextual reasoning, principled decision-making, and core components of high-quality lesson planning (Guo & Lee, 2023). Together, these contrasting perspectives highlight the need to investigate not only the efficiencies introduced by AI but also the degree to which such tools genuinely deepen teachers' analytical and creative engagement during planning.

The cognitive demands of lesson planning are particularly salient in EFL settings, where teachers must evaluate linguistic goals, anticipate learner needs, and integrate pedagogical decisions within curriculum constraints. Frameworks such as cognitive apprenticeship and transfer of learning offer useful lenses for understanding how teachers engage with support tools during this process. The former highlights the value of guided modeling and scaffolded reasoning in complex tasks (Collins et al., 1989), while the latter explains how strategies developed under supportive conditions may- or may not- transfer to independent performance (Perkins & Salomon, 1992). These perspectives suggest that interactions with AI could either strengthen teachers' internal reasoning or encourage superficial choices that mask gaps in analytical and creative thinking, making it essential to investigate how such tools shape cognition in authentic planning contexts.

Despite growing interest in AI-supported pedagogy, significant gaps remain in understanding how such tools influence teachers' cognitive work during lesson planning. Much of the existing literature concentrates on student outcomes, leaving limited evidence on how AI shapes teachers' analytical reasoning, creative decision-making, and capacity to justify instructional choices. Empirical studies involving in-service EFL teachers are relatively scarce, and even fewer examine planning tasks embedded in real professional contexts where judgments must align with curricular, linguistic, and pedagogical demands. Moreover, research rarely integrates theoretical perspectives with rigorous designs capable of linking observable improvements to underlying cognitive processes, creating a need for studies that combine measurable outcomes with teachers' own accounts of how AI mediates their planning. By focusing on in-service EFL teachers engaged in real lesson-planning tasks and combining measurable outcomes with first-hand cognitive accounts, the present study extends prior work that has largely centered on students or survey-based evaluations.

To address these gaps, the present study examined how ChatGPT-assisted lesson planning shaped EFL teachers' critical thinking, creativity, and the overall quality of the lesson plans they produced. Employing an explanatory sequential mixed-methods design with experimental and control groups, the study integrated quantitative evidence with teachers' qualitative accounts to capture both measurable outcomes and the cognitive processes that accompanied planning with AI support. By combining these complementary perspectives, the research aimed to provide a more balanced understanding of the pedagogical affordances and limitations of ChatGPT in professional teaching contexts. Accordingly, the study addressed the following questions:

1. Does ChatGPT-assisted lesson planning influence EFL teachers' critical thinking, creativity, and overall lesson-plan quality compared with conventional planning practices?
2. How do teachers perceive and experience the use of ChatGPT during lesson planning, particularly with respect to their cognitive engagement, decision-making processes, and pedagogical reasoning?
3. What benefits and limitations do teachers identify regarding ChatGPT's role in supporting efficiency, originality, and professional autonomy in lesson planning?

LITERATURE REVIEW

COGNITIVE FOUNDATIONS OF LESSON PLANNING AND AI-SUPPORTED PEDAGOGY

Lesson planning is a cognitively demanding process requiring teachers to evaluate learner needs, select appropriate instructional approaches, and construct coherent sequences of activities. Two cognitive capacities- critical thinking and creativity- play a central role in this work. Critical thinking enables teachers to analyze assumptions, assess evidence, and justify pedagogical choices (Paul & Elder, 2019), while creativity supports the generation of original ideas and the adaptation of instructional content to students' linguistic and cultural contexts (Sawyer, 2011; Richards, 2013). Within EFL settings, these complementary capacities shape teachers' ability to design lessons that are both pedagogically sound and responsive to classroom realities.

Several theoretical traditions help explain how teachers develop these capacities during complex planning tasks. Cognitive apprenticeship emphasizes guided modeling, scaffolding, and gradual independence as teachers engage with authentic problem-solving (Collins et al., 1989). Similarly, transfer of learning theory highlights how strategies acquired in supported environments may be internalized and used independently (Perkins & Salomon, 1992). These frameworks offer valuable lenses for examining how teachers interact with emerging AI tools during lesson planning: while structured guidance from AI may help teachers explore new possibilities, it may also lead to surface-level performance if users rely on suggestions without engaging in deeper reasoning. Seen through these perspectives, understanding teachers' cognitive processes-

not merely the outputs they produce- becomes essential in evaluating the pedagogical value of generative AI.

AI-ASSISTED LESSON PLANNING: OPPORTUNITIES AND RISKS

The emergence of tools such as ChatGPT has transformed how teachers approach instructional planning. AI systems can generate initial ideas, reorganize content, and propose alternative pathways for presenting concepts, potentially expanding teachers' repertoires and reducing preparation time (Karakaş & Yesilyurt, 2023). Recent research from diverse educational contexts suggests that generative AI may stimulate creative ideation and broaden teachers' exposure to instructional models. For instance, Kartal (2024) found that iterative interactions with ChatGPT supported EFL student teachers' creative thinking and encouraged more imaginative task design. Likewise, Liu et al. (2023) reported that teachers using AI exhibited more divergent thinking and greater flexibility during planning.

A growing body of work also highlights AI's role in supporting professional development. European Schoolnet initiatives document how AI-driven tools can enhance teachers' reflective engagement and instructional decision-making when used intentionally (Cukurova et al., 2024). Studies in technology-enhanced education further demonstrate how AI can assist teachers in personalizing instruction and responding more quickly to curriculum demands (Do Amaral, 2024; Wu, 2024).

However, the pedagogical benefits of AI-assisted planning are accompanied by important concerns. Research on cognitive offloading suggests that heavy reliance on AI tools may weaken opportunities for independent reasoning and justification. Dahri et al. (2024) observed that teachers who used ChatGPT uncritically often produced lesson plans with limited analytical depth. Similarly, De La Puente et al. (2024) found that AI-generated feedback sometimes constrained reflective thinking by offering prescriptive solutions that users accepted without questioning. Guo and Lee (2023) also warned that habitual use of AI tools may narrow teachers' problem-solving skills by prioritizing convenience over cognitive engagement.

Empirical studies at the school level reflect the same tension. Powell and Courchesne (2024), for example, showed that ChatGPT-generated lesson plans for primary science classes provided efficient structures but occasionally failed to account for curricular nuances and learner variability. Such findings underscore the need for teachers to maintain evaluative and adaptive engagement when using AI, rather than merely imitating AI-generated suggestions.

EVIDENCE FROM EFL AND IN-SERVICE TEACHING CONTEXTS

Despite increasing interest in AI-supported pedagogy, research involving in-service EFL teachers remains limited. Much of the existing work focuses on pre-service teachers (Karakaş & Yesilyurt, 2023; Kartal, 2024; Pişkin Tunç, 2024) or students in higher education contexts (Toma & Pérez, 2024; Zhua et al., 2024). While these studies provide valuable insights into the potential benefits of AI, they offer relatively little information about how practicing teachers engage cognitively with AI during real lesson-planning tasks that require balancing curriculum demands, linguistic goals, and classroom constraints.

A review by Ng et al. (2023) highlights this gap, noting that research on AI in education has tended to emphasize student learning outcomes rather than teachers' professional reasoning or cognitive workload. Studies that do involve teachers often rely on self-reported perceptions rather than examining actual planning processes or evaluating the quality of the resulting lesson plans (Akrami & Ghaderi, 2024; Den Berg & Du Plessis, 2023). Even fewer studies link teachers' use of AI to measurable changes in critical thinking or creativity.

Emerging research specifically addressing ChatGPT's role in lesson planning offers mixed findings. Yildiz (2024) reported that EFL teachers gained confidence and demonstrated improved planning skills when using AI as a support tool. In contrast, Stognieva (2024) found that AI-generated suggestions sometimes encouraged overly standardized lesson structures that limited originality. These contrasting results illustrate the complex and context-dependent nature of AI's influence on teacher cognition.

Across the reviewed literature, the key theoretical, methodological, and contextual limitations remain closely connected and continue to restrict a clear understanding of how AI-supported lesson planning shapes teachers' cognitive development. Theoretically, few studies draw on frameworks such as cognitive apprenticeship or transfer of learning, leaving uncertain whether AI fosters deeper reasoning or simply increases efficiency. Methodologically, the dominance of descriptive and survey-based designs offers limited insight into causal mechanisms, highlighting the need for mixed-methods and experimental research capable of capturing both outcomes and underlying reasoning processes.

Contextually, evidence from authentic in-service EFL settings is scarce, particularly studies examining real planning tasks aligned with curricular demands. This gap limits our understanding of how AI affects teachers' professional autonomy, analytical reasoning, and creative judgment, and prevents the field from forming a comprehensive account of AI's impact on practicing language teachers.

METHODOLOGY

RESEARCH DESIGN

This study employed an explanatory sequential mixed-methods design embedded within a quasi-experimental pretest–posttest control-group structure. The quantitative strand constituted the primary component of the investigation and examined whether sustained engagement with ChatGPT during lesson planning influenced teachers' critical thinking, creativity, and overall lesson-plan quality. The qualitative strand followed the statistical analyses and sought to clarify how participants experienced the intervention, what forms of reasoning they applied while

interacting with the AI tool, and how these insights could help interpret the quantitative patterns. This methodological configuration aligns with established mixed-methods traditions that emphasize the integration of numerical trends with contextualized participant accounts in technology-enhanced educational research (Creswell & Plano Clark, 2018).

The choice of an explanatory sequential structure was informed by both methodological priorities and the nature of the research questions. Prior work in AI-supported pedagogy has shown that quantitative gains alone cannot fully reveal the cognitive processes teachers engage in when working with generative language models. Documenting measurable outcomes in the first phase and subsequently pursuing interpretive explanations through qualitative inquiry enabled a more comprehensive understanding of how AI-assisted planning shaped professional reasoning. In addition, implementing a quasi-experimental arrangement with two parallel groups- a treatment group (n = 21) receiving structured ChatGPT-based training and a control group (n = 21) engaged in conventional non-AI lesson planning- strengthened the study's internal validity by supporting direct comparisons of changes over time under clearly differentiated instructional conditions.

The design and implementation of the intervention were grounded in two complementary theoretical perspectives: the cognitive apprenticeship framework (Collins et al., 1989) and the theory of transfer of learning (Perkins & Salomon, 1992). As presented in the Literature Review, the cognitive apprenticeship model emphasizes guided modeling, scaffolding, and the gradual release of responsibility- principles that were operationalized through structured prompting, reflective analysis, and iterative revision cycles incorporated into the ChatGPT-assisted planning sessions for the experimental group. The transfer-of-learning perspective further posits that strategies internalized during supported practice can generalize to contexts where such support is absent. Consistent with this idea, the posttest required all 42 participants- both the experimental and control groups- to produce a lesson plan without access to ChatGPT. This design enabled the study to assess whether cognitive or pedagogical benefits acquired during the intervention persisted beyond the period of AI support. The integration of these theoretical lenses provided a coherent foundation for the structure of the training sequence and informed the interpretation of both quantitative and qualitative findings.

PARTICIPANTS

Participants were selected from 73 eligible in-service EFL teachers employed in private language institutes in Shahrekord (Safir Language Academy). Selection was conducted through institute-wide announcements and local professional teacher networks. Eligibility criteria required: (a) two to fifteen years of EFL teaching experience, (b) age between 20 and 40 years, (c) prior completion of at least one formal lesson-planning or teacher-training course, and (d) full availability to attend all intervention sessions. They were chosen because they represent the typical recruitment profile of private EFL institutes in Iran and ensure participants possess sufficient professional maturity while avoiding late-career specialization effects. Thirty-one applicants were excluded prior to baseline due to ineligibility (n = 18), voluntary withdrawal before pretest administration (n = 8), or absence from the mandatory orientation session (n = 5). The final sample consisted of 42 teachers who provided written informed consent and completed the baseline assessment. Ethical approval for the study was granted by the Institutional Review Board of Safir Language Academy. Descriptive characteristics

Table 1. Demographic Characteristics of Participants (N = 42)

Variable	Experimental Group (n = 21)	Control Group (n = 21)	Total
Age (years)	M = 29.2 (SD = 4.3)	M = 29.6 (SD = 4.1)	Range: 20–40
Gender	12 Female / 9 Male	11 Female / 10 Male	23 F / 19 M
Teaching Experience (years)	M = 7.0 (SD = 3.1)	M = 7.2 (SD = 3.3)	Range: 2–15
English Proficiency Level	Intermediate: 12 Upper-intermediate: 9	Intermediate: 13 Upper-intermediate: 8	—
Prior Technology Use	Low: 4 Moderate: 11 High: 6	Low: 5 Moderate: 10 High: 6	—
Prior Training in Lesson Planning	Yes: 21	Yes: 21	42

Note. Across both groups, 25 teachers were categorized as intermediate and 17 as upper-intermediate. For technology use, 9 reported low use, 21 moderate use, and 12 high use.

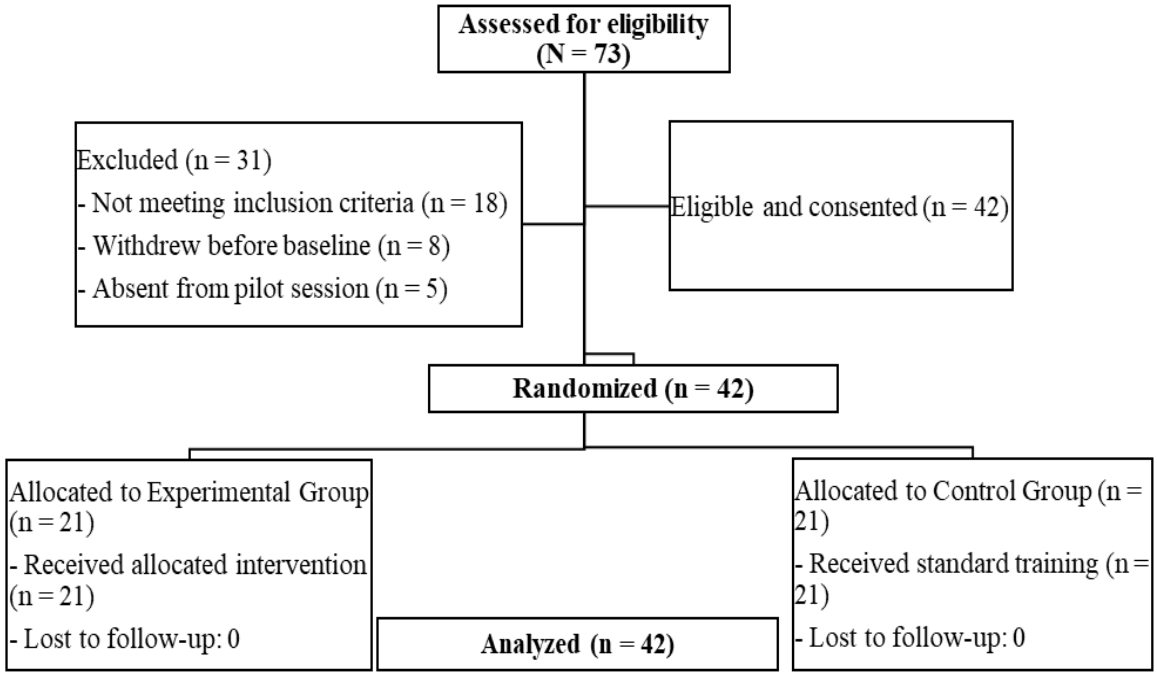

Figure 1. CONSORT-style flow diagram of participant recruitment and allocation.

(age, gender, years of teaching, proficiency band, and prior training) are summarized in Table 1.

To establish baseline equivalence between groups, all consenting participants completed an institutional CEFR-aligned English proficiency diagnostic test and an 8-item technology-use questionnaire whose content validity was confirmed by two TEFL specialists. These preliminary scores informed a stratified random assignment procedure in which participants were allocated, via computer-generated random numbers, to either the experimental group (n = 21), which received structured ChatGPT-assisted lesson-planning training, or the control group (n = 21), which continued using conventional non-AI planning procedures. This assignment process produced two groups comparable in English proficiency, teaching experience, and prior technology familiarity (Figure 1).

A detailed overview of the recruitment, screening, and allocation process is illustrated in the CONSORT-style flow diagram.

INSTRUMENTS

To evaluate teachers' cognitive and pedagogical performance through their lesson plans, three complementary instruments were employed: an analytic rubric for assessing critical thinking, an analytic rubric for evaluating creativity, and a holistic scale measuring overall lesson-plan quality. The analytic rubrics were adapted from the Association of American Colleges and Universities (AAC & U VALUE frameworks, 2009) and revised to align with the specific cognitive and pedagogical demands of EFL lesson planning. Adaptation procedures included contextualizing the original descriptors for ELT instructional decision-making, refining analytical categories to reflect the realities of lesson preparation, and adjusting the performance levels to capture qualitative distinctions typically observed in teachers' planning processes. Full versions of the adapted rubrics appear in Appendix A (Critical Thinking Rubric) and Appendix B (Creativity Rubric).

To establish the content validity of these adapted instruments, the rubrics were reviewed by three PhD-holding experts in TEFL and teacher education, each with over ten years of experience in language assessment and curriculum design, who evaluated their relevance, clarity, and consistency with current ELT pedagogical standards. Based on their suggestions, several descriptors were reordered to improve precision. A brief pilot application of the rubrics to four sample lesson plans was also conducted to confirm the interpretability and practical usability of the criteria before formal scoring began.

The Critical Thinking Rubric evaluates five dimensions central to pedagogical reasoning: (a) identification and clarification of the instructional problem, (b) justification of instructional decisions using pedagogical evidence, (c) recognition of underlying assumptions, (d) consideration of alternative instructional options, and (e) articulation of coherent instructional conclusions. The Creativity Rubric assesses originality, flexibility in planning, innovative task design, productive risk-taking, and the integration of diverse instructional approaches. Each rubric contains five criteria rated on a four-point scale (1 = Emerging to 4 = Exemplary), generating total domain scores ranging from 5 to 20.

In addition to these analytic measures, a holistic Lesson Plan Quality Score (LPQS) was developed to provide a broader evaluation of overall instructional quality. The LPQS criteria address lesson coherence and sequencing, alignment between objectives and activities, contextual appropriateness of pedagogical choices, methodological soundness, and clarity of instructional procedures. This

scale offers a global assessment that complements the finer-grained evaluations of the analytic rubrics. Full LPQS criteria and scoring guidelines are included in Appendix C. These instruments were applied uniformly to all lesson plans produced in the pretest and posttest phases by both the experimental and control groups.

PROCEDURE

The intervention was implemented over an eight-week period and followed a structured sequence consisting of an orientation stage, a pretest, a four-week instructional intervention, a posttest, and follow-up qualitative interviews. All sessions were delivered in face-to-face format at a centralized teacher-training facility affiliated with the participating language institutes. Each weekly session lasted approximately two hours.

Before beginning formal data collection, orientation and pilot session was run. All participants attended a mandatory orientation session in which the study procedures, assessment requirements, and ethical guidelines were explained. During this session, four pilot lesson plans (unrelated to the study tasks) were jointly reviewed to ensure that participants understood the scoring criteria and expectations of the planning tasks.

At the first week, the pretest phase was done. In this time, both the experimental group (n = 21) and the control group (n = 21) completed an identical pretest task requiring them to design a lesson plan without access to any AI tools. These plans were evaluated using the Critical Thinking Rubric, Creativity Rubric, and LPQS to establish baseline cognitive and pedagogical performance.

During weeks 2–5 which is called the intervention phase, the two groups followed distinct procedures. In the experimental Group (ChatGPT-Assisted Planning), participants received structured training on integrating ChatGPT into lesson-planning processes. Each weekly session included (a) a short modeling segment in which the facilitator demonstrated targeted prompting strategies (e.g., task-specific prompting, refinement prompts, and idea-generation prompts), (b) guided planning cycles in which teachers interacted with ChatGPT to generate, evaluate, and revise lesson components, and (c) reflection activities in which participants justified their instructional decisions and evaluated the reliability and pedagogical value of AI-generated content. In the control Group (Conventional Planning), participants completed the same weekly planning tasks but relied solely on traditional methods such as textbooks, syllabi, and their prior pedagogical knowledge. They received no exposure to AI tools and followed standard institute procedures for lesson planning. Both groups produced one complete lesson plan per week, ensuring parallel task exposure and equivalent cognitive demands across conditions.

In Week 6, which is called posttest phase, all participants completed a new lesson-planning task without access to ChatGPT. This requirement was included to examine whether the strategies practiced during the intervention transferred to an AI-free context. The same assessment instruments used at pretest were applied to evaluate the posttest lesson plans. Week 7 was reserved for delayed submissions and clarification sessions to ensure full data completion.

In the final week, semi-structured interviews were conducted with 12 participants- six from the experimental group and six from the control group- selected using maximum-variation sampling based on years of teaching experience and pretest/posttest performance patterns (Appendix D). The interviews explored participants' planning experiences, cognitive processes, perceived affordances and limitations of AI support, and reflections on the instructional tasks. Each interview lasted approximately 25-35 minutes and was audio-recorded with participants' consent.

DATA ANALYSIS

QUANTITATIVE ANALYSIS

The quantitative phase employed a repeated-measures multivariate analysis of variance (MANOVA) to evaluate the effects of the intervention on three interrelated dependent variables - critical thinking, creativity, and overall lesson-plan quality- across two time points (pretest, posttest) and two groups (experimental vs. control). MANOVA was selected because it permits simultaneous testing of mean differences on correlated outcome measures while controlling familywise error and enhancing statistical power relative to separate univariate tests.

Prior to conducting MANOVA, an a priori statistical power analysis (G*Power) was performed, indicating that a sample size of approximately 40 participants would be sufficient to detect medium-sized effects (Cohen's $f \approx .25$) with 80% power in a repeated-measures design. The final sample (n = 42) met this threshold. Assumption checks included examinations of multivariate normality (Shapiro–Wilk on residuals and visual inspection of Q-Q plots and histograms), equality of covariance matrices (Box's M test), and homogeneity of error variances (Levene's test for each dependent variable). Given the two-time-point repeated structure, sphericity was not applicable; nevertheless, where appropriate, robust procedures (reporting Pillai's Trace) and bootstrapped estimates were planned as remedies for any assumption violations.

The primary multivariate test reported will be Pillai's Trace (supplemented by Wilks' Lambda for completeness). If the MANOVA indicated a statistically significant multivariate effect for time, group, or their interaction, follow-up analyses would proceed in a hierarchical manner: (a) univariate repeated-measures ANOVAs for each dependent variable, (b) effect sizes (partial η^2) reported for all significant effects, and (c) pairwise comparisons with multiplicity correction (Holm procedure or Bonferroni, based on the number of comparisons) to identify the locus of change. All reported p-values will be two-tailed, and 95% confidence intervals will accompany estimates of means and effect sizes. Descriptive statistics (means and standard deviations) for each dependent variable by group and time will be presented in the Results section.

Missing data and outliers were addressed according to pre-established rules. Data were inspected for missingness patterns; if missing values were below 5% and missing completely at random (MCAR), listwise deletion was applied for the affected analyses. If missingness exceeded this threshold or suggested non-random patterns, multiple imputation (MI; m = 20 imputations) was planned and results compared to complete-case analyses as a sensitivity check. Outliers were screened using standardized scores (±3 SD) and influence diagnostics; sensitivity analyses excluding extreme cases were performed when appropriate. All quantitative analyses were executed using SPSS (version 28) and supplemented with R (version 4.x) for diagnostic plots and bootstrapped confidence intervals. G*Power (version 3.1) was used for power analysis. Exact analytic scripts and output files are available upon request.

QUALITATIVE ANALYSIS

Interview data were analyzed using reflexive thematic analysis following Braun and Clarke's (2006) six-phase procedure: (1) familiarization with the data, (2) generation of initial codes, (3) searching for candidate themes, (4) reviewing and refining themes, (5) defining and naming themes, and (6) producing the report. Transcripts were anonymized and imported into NVivo 12 for coding and data management.

To enhance analytical rigor, a subset of interview transcripts (30% of the sample) was double-coded independently by two experienced coders; inter-coder agreement was quantified using Cohen's Kappa ($\kappa = 0.81$), indicating substantial agreement (McHugh, 2012). Coding discrepancies were resolved through discussion and consensus; an audit trail documenting coding decisions and codebook revisions was maintained. Data collection ceased when thematic saturation was achieved (i.e., no new themes emerged across successive interviews), and this saturation point is reported in the Results. Member checking (brief verification of interpretive summaries with selected participants) and reflexive memoing were used to support credibility and confirmability (also see Appendix D).

INTEGRATION STRATEGY (MIXED-METHODS SYNTHESIS)

Integration of quantitative and qualitative evidence occurred at two stages. First, after completing within- and between-group quantitative analyses, we constructed joint displays (tables combining group-level changes in mean scores with representative qualitative themes) to examine convergence, complementarity, and divergence between data strands. Second, qualitative findings were used to interpret and explain specific quantitative patterns (e.g., why certain participants showed greater gains on creativity but not on critical thinking). Illustrative joint displays and the final thematic table (themes, subthemes, exemplar quotes) are presented in Appendix E.

All statistical tests, effect sizes, and confidence intervals will be reported transparently. Following best practices for reproducible research, analytic code, de-identified quantitative datasets, and coding frameworks for qualitative analysis are archived and will be made available on reasonable request, subject to ethical and privacy constraints. Analyses were performed with SPSS v28, R v4.x, NVivo 12, and G*Power 3.1. The study protocol, including analytic plans, was registered prior to data analysis (registration details provided in Appendix E). Ethical approvals and informed consent procedures are described in the Methods section.

RESULTS

PRELIMINARY ANALYSES AND DESCRIPTIVE STATISTICS

The analysis began with an examination of descriptive statistics for the three outcome variables, critical thinking, creativity, and overall lesson-plan quality, across the two testing points and experimental conditions. At the pretest stage, the experimental and control groups demonstrated comparable mean scores on all three measures, suggesting baseline equivalence prior to the intervention. Minor variations in central tendencies and dispersion were observed; however, independent-samples t-tests conducted on the pretest data confirmed that none of these differences reached statistical significance. This initial comparability ensured that any subsequent divergence between groups could be more confidently attributed to the ChatGPT-assisted training rather than pre-existing disparities. Full descriptive statistics, including mean scores, standard deviations, and confidence intervals for each measure, are

Table 2. Descriptive Statistics for Outcome Measures by Group and Time (Means and Standard Deviations)

Outcome Measure	Group	Pretest M (SD)	Posttest M (SD)
Critical Thinking	Experimental	12.38 (2.11)	15.76 (2.34)
	Control	12.19 (2.04)	13.02 (2.15)
Creativity	Experimental	12.90 (2.25)	16.10 (2.41)
	Control	12.72 (2.18)	13.48 (2.26)
Lesson-Plan Quality	Experimental	13.05 (2.31)	16.48 (2.53)
	Control	12.90 (2.27)	13.85 (2.32)

Table 3. Summary of Follow-up Univariate ANOVAs for Outcome Measures

Outcome Variable	Effect	F(1, 40)	p	Partial η²
Critical Thinking	Time	29.12	< .001	.42
	Group	1.02	.32	.03
	Time × Group	18.44	< .001	.32
Creativity	Time	24.51	< .001	.38
	Group	0.76	.39	.02
	Time × Group	16.87	< .001	.30
Lesson-Plan Quality	Time	33.45	< .001	.46
	Group	1.18	.28	.03
	Time × Group	20.33	< .001	.34

Note. All ANOVAs were conducted following a significant multivariate interaction effect reported in Section 3.2. Effect sizes (partial η²) indicate medium-to-large magnitudes across all three outcomes, consistent with the intervention's expected impact based on descriptive trends (Table 2).

presented in Table 2.

Initial analysis confirmed baseline equivalence across all three outcomes (Table 2). Post-intervention, the experimental group showed notable gains in critical thinking, creativity, and lesson-plan quality, while the control group showed minimal improvement. To determine if these observed differences represent statistically significant effects rather than random variance, a repeated-measures MANOVA was employed. To evaluate the overall impact of the intervention on the three outcome measures, a repeated-measures MANOVA was conducted with time (pretest vs. posttest) as the within-subject factor and group (experimental vs. control) as the between-subject factor. The multivariate test yielded a significant time × group interaction (Pillai's Trace = .41, $F(3, 38) = 8.87$, $p < .001$, partial $\eta^2 = .41$), indicating that the pattern of change across the dependent variables differed markedly between the two groups. The main effect of time was also significant (Pillai's Trace = .56, $F(3, 38) = 16.36$, $p < .001$, partial $\eta^2 = .56$), suggesting substantial improvement across participants from pretest to posttest. In contrast, the multivariate main effect of group was non-significant (Pillai's Trace = .07, $F(3, 38) = 0.96$, $p = .42$), partial $\eta^2 = .07$, confirming that no meaningful differences existed between the experimental and control groups prior to the intervention. Taken together, these findings demonstrate that the ChatGPT-assisted training produced differential gains over time, with the experimental group showing significantly greater improvement across all three cognitive and pedagogical dimensions than the control group. Follow-up repeated-measures ANOVAs revealed significant time × group interactions for critical thinking ($F(1, 40) = 18.44$, $p < .001$, partial $\eta^2 = .32$), creativity ($F(1, 40) = 16.87$, $p < .001$, partial $\eta^2 = .30$), and lesson-plan quality ($F(1, 40) = 20.33$, $p < .001$, partial $\eta^2 = .34$). These interactions indicate substantially greater pretest-to-posttest score increases in the experimental group compared to the control group. Significant main effects for time were observed across all variables (ps < .001), confirming temporal changes, while non-significant main effects for group (ps > .28) corroborated comparable baseline performance. These findings demonstrate consistent, domain-wide advantages with medium-to-large effect sizes for the ChatGPT-assisted intervention.

EFFECT SIZE INTERPRETATION AND MAGNITUDE OF CHANGE

To contextualize the statistical results, the effect sizes associated with the follow-up ANOVAs were examined to determine the practical significance of the observed improvements. Across all three outcome measures, the interaction effects demonstrated medium-to-large magnitudes (partial η^2 ranging from .30 to .34), indicating that the gains made by the experimental group were not only statistically significant but also educationally meaningful. These values suggest that a substantial proportion of the variance in posttest performance can be attributed to the ChatGPT-assisted training rather than general developmental change or repeated exposure to the assessment tasks. The main effects of time also showed large effect sizes (partial $\eta^2 = .38$ to .46), reflecting overall growth in lesson-planning competence across participants. In contrast, the negligible effect sizes for the main effects of group at pretest (partial $\eta^2 = .02$ to .03) confirm that the two groups entered the intervention with comparable initial abilities. Collectively, these patterns indicate that the ChatGPT-supported instructional approach had a robust and domain-wide impact on teachers' cognitive and pedagogical performance.

QUALITATIVE FINDINGS

Semi-structured interviews were conducted with a purposive subsample of 12 participants, including six teachers from the experimental group and six from the control group, to further illuminate the cognitive and pedagogical processes underlying the quantitative results. Thematic analysis followed Braun and Clarke's (2006) six-phase framework, supported by systematic coding procedures documented in Appendix D. Three

overarching themes and eight subthemes were identified through iterative review and refinement of the coded data. These themes capture participants' experiences with lesson planning, their cognitive engagement during the intervention, and their perceptions of the affordances and constraints of AI-based support. For instance, some teachers reported that ChatGPT allowed them to explore new and diverse approaches to lesson design, particularly in creating engaging activities and instructional strategies. One teacher noted, "ChatGPT opened my eyes to new ways of making lessons more interactive. It gave me a variety of ideas that I wouldn't have thought of on my own." Another participant mentioned, "The flexibility it provides is invaluable. I can adapt lesson plans quickly based on students' needs, which I couldn't do as easily before." While seven teachers appreciated the benefits of AI in lesson planning, nine expressed concerns about the alignment of AI-generated ideas with their personal teaching styles. Other participants from the experimental group indicated that while ChatGPT helped generate ideas, they often had to modify the content to fit their specific classroom context. One teacher shared, "I can see how ChatGPT is helpful, but sometimes the suggestions feel too generic. I have to spend extra time tweaking them to make them suitable for my students' level." Another remarked, "The ideas it suggests are great, but not always aligned with my teaching philosophy. I prefer more structured lessons, and sometimes AI offers too much variety."

Accordingly, representative excerpts illustrating each theme are presented in Appendix E. The following sections outline the major themes that emerged from the analysis.

Theme 1- Enhanced Cognitive Structuring Through AI-Supported Planning: participants in the experimental group consistently described the AI-assisted planning sessions as a catalyst for organizing and deepening their instructional reasoning. Teachers reported that interacting with ChatGPT prompted them to articulate pedagogical decisions more explicitly, compare alternative instructional pathways, and justify their choices with greater clarity. This structured reasoning was perceived as particularly valuable during the early stages of the planning process, where the AI's probes and reformulations helped participants refine lesson objectives, anticipate learner needs, and consider multiple pedagogical routes before committing to a final plan. By contrast, teachers in the control group more often characterized their planning as "linear" or "habit-driven," indicating fewer opportunities for deliberate cognitive exploration. As one participant in the experimental group explained, "When ChatGPT asked me why I chose a particular activity, it made me stop and think. I had to evaluate my reasoning, not just follow my usual routine." Such accounts align with the quantitative gains observed in critical thinking and suggest that AI-supported planning promoted a more reflective and analytically structured approach to instructional design.

Theme 2- Expansion of Creative Pedagogical Repertoires Through Iterative AI Interaction: in contrast, control-group teachers more commonly reported relying on familiar templates and established lesson structures, often citing time constraints and limited exposure to new pedagogical models. The divergence between groups mirrors the significant quantitative gains observed in creativity scores, suggesting that sustained engagement with AI facilitated an expanded creative repertoire. As one experimental participant noted, "The ideas from ChatGPT pushed me to think differently- not to copy, but to adapt and combine. It helped me approach planning with more openness and imagination." These patterns indicate that the AI-supported design environment functioned as a generative space in which teachers could explore alternatives, test possibilities, and cultivate more innovative lesson-planning habits.

Theme 3- Sustained Instructional Gains Beyond AI Assistance: a notable finding from the qualitative phase was the extent to which several experimental-group teachers reported applying the reasoning strategies they had practiced with ChatGPT even after the tool was removed in the posttest task. Participants described continuing to use habits formed during the intervention- such as systematically examining alternative instructional options, articulating rationales for each planning decision, and anticipating learner responses. These self-directed cognitive routines suggest that aspects of the AI-supported process were internalized and carried over into independent practice, consistent with the theoretical expectations of transfer of learning.

Teachers emphasized that the AI's questioning patterns and scaffolding during the training phase had prompted them to adopt a more deliberate, inquiry-oriented approach to planning. During interviews, some explained that although the posttest required them to work without ChatGPT, they found themselves "mentally rehearsing" the kinds of prompts, comparisons, and clarifications that had characterized the earlier AI-assisted sessions. Such reflections provide qualitative support for the significant posttest improvements observed in both critical thinking and overall lesson-plan quality within the experimental group.

By contrast, teachers in the control group generally reported less change in their independent planning routines. Their descriptions often reflected continuity rather than transformation, with most indicating they followed familiar procedures similar to those used in the pretest. These contrasting patterns underscore the role of structured AI-supported scaffolding in fostering transferable cognitive strategies that persist beyond the immediate intervention environment.

Theme 4- Differential Cognitive Benefits Across Domains: interview data revealed that the cognitive outcomes of the intervention did not unfold uniformly across the three focal domains- critical thinking, creativity, and overall lesson-plan quality. Participants in the experimental group generally reported that ChatGPT was particularly effective in stimulating idea generation, re-framing instructional problems, and proposing alternative task designs. These affordances appeared to support creativity more directly than critical thinking, offering teachers a broader repertoire

of instructional options from which to build and refine their plans.

Teachers widely agreed that ChatGPT's greatest contribution lay in expanding the range of instructional possibilities they could consider. However, five of the eight interviewed teachers in the experimental group reported that although ChatGPT consistently generated diverse and original options, these suggestions did not automatically translate into deeper evaluative reasoning. They explained that the tool excelled at producing variation- alternative task designs, different sequencing options, or multiple ways of framing instructional goals- but the responsibility for determining which choices were pedagogically justified ultimately remained with the teacher. This pattern aligns with the quantitative findings, in which gains in creativity were more pronounced and consistent across participants, whereas gains in critical thinking showed greater individual variability.

Teachers who exhibited stronger posttest improvements in critical thinking described a more reflective interaction style during the AI-assisted phase. They frequently paused to interrogate the tool's suggestions, compared alternatives, and asked follow-up questions aimed at clarifying assumptions or instructional consequences. By contrast, teachers who used ChatGPT more passively- accepting suggestions at face value- tended to show more modest critical thinking gains. These accounts illuminate the mechanisms underlying the quantitative results, particularly the smaller effect sizes observed for critical thinking relative to creativity.

The control group displayed a more uniform pattern: most participants followed established planning routines, resulting in limited variation between pretest and posttest performance. Their interview reflections suggested incremental refinements rather than substantial shifts in conceptual or procedural thinking. This contrast reinforces the interpretation that the structured, reflective use of ChatGPT was essential to the differential cognitive development observed in the experimental group.

Theme 5- Shifting Perceptions of AI's Role in Professional Practice: a notable shift emerged in how teachers in the experimental group conceptualized the role of AI in their professional practice. Prior to the intervention, many participants viewed AI as a peripheral or optional tool- useful primarily for generating quick ideas or simplifying routine tasks. By the end of the training, however, 15 out of 21 experimental participants described AI as a "collaborative thinking partner," emphasizing its capacity to support reflection, extend cognitive reach, and stimulate more deliberate instructional decision-making. Rather than perceiving ChatGPT as a shortcut, teachers increasingly framed it as a resource for enhancing planning depth, structuring pedagogical reasoning, and expanding instructional possibilities. Importantly, several teachers highlighted that the value of AI was maximized not when they accepted its suggestions as-is, but when they critically evaluated, adapted, and integrated those ideas into contextually responsive lesson plans. This perspective aligns with the measurable improvement observed in critical thinking scores for the experimental group.

In contrast, teachers in the control group-who planned without AI- tended to conceptualize lesson planning as a manual, time-intensive process grounded primarily in personal experience and established routines. While they acknowledged the potential usefulness of AI in general terms, only 4 out of 21 control participants reported that they would consider incorporating such tools into future practice. Their reservations largely stemmed from concerns about accuracy, overreliance, and diminishing teacher autonomy- concerns that experimental-group participants also voiced at the beginning of the study but reported overcoming as they gained structured experience. The contrast between the two groups suggests that hands-on, guided interaction with AI plays a decisive role in shaping professional perceptions and fostering more informed, balanced understandings of its pedagogical affordances and limitations.

INTERPRETATION OF QUANTITATIVE AND QUALITATIVE FINDINGS

The integration of quantitative and qualitative results provides a consolidated understanding of how the ChatGPT-assisted intervention influenced teachers' cognitive and pedagogical performance. Quantitatively, the experimental group demonstrated significantly greater gains than the control group across all three outcomes, domains- critical thinking, creativity, and overall lesson-plan quality- indicating that the intervention generated measurable improvements in teachers' planning competencies. Qualitatively, the five themes revealed specific mechanisms that help explain these statistical patterns, including heightened cognitive engagement, expansion of creative pedagogical repertoires, and more reflective, evidence-based reasoning during the planning process. When interpreted together, the two strands suggest that the observed quantitative gains were not merely numerical changes but reflected deeper shifts in teachers' instructional thinking, the range of strategies they considered, and the ways they approached planning tasks with and without AI support.

A strong agreement was found between the quantitative results and qualitative themes. The experimental group's significant rise in creativity scores mirrored teachers' self-reports of enhanced idea generation, adaptive task management, and innovative instructional approaches during AI-assisted planning. Likewise, critical thinking enhancements correlated with qualitative descriptions of more reasoned pedagogical decision-making, thorough assumption analysis, and explicit articulation of teaching strategies. The qualitative data also enriched the quantitative findings by explaining the underlying mechanisms, such as iterative AI prompting, reflective consideration of AI suggestions, and heightened self-awareness during planning. Minor discrepancies were noted, with a few participants finding it challenging to maintain the same level of cognitive engagement in independent planning post-intervention, indicating potential individual variability

Table 4. Joint Display Integrating Quantitative Outcomes and Qualitative Themes (RQ1-RQ3)

Quantitative Findings	Corresponding Qualitative Themes	Integrated Interpretation (Meta-Inference)	Implications
Significant time × group interaction across all three domains: critical thinking, creativity, and lesson-plan quality (Pillai's Trace = .41, $p < .001$).	Theme 1: Heightened Cognitive Engagement in AI-Supported Planning	AI-supported prompting increased depth of reasoning, clarification of instructional choices, and metacognitive awareness, explaining stronger gains.	Guided AI interaction can enhance cognitive processing in teacher education.
Largest effect size observed for creativity ($\eta^2 = .32$).	Theme 2: Expansion of Creative Pedagogical Repertoires	Teachers used ChatGPT suggestions as starting points for innovative adaptations, producing broader and more flexible designs.	AI can effectively function as a creativity amplifier in instructional design.
Significant improvement in critical thinking scores for the experimental group ($\eta^2 = .27$).	Theme 3: Strengthening of Evidence-Based Pedagogical Reasoning	Participants justified decisions more rigorously, compared options, and examined assumptions—behaviors mirrored in quantitative gains.	Integrating AI can scaffold deeper pedagogical reasoning.
Posttest performance remained higher even without AI access.	Theme 4: Internalization and Transfer of Planning Strategies	Teachers applied AI-supported strategies (evaluation, justification, sequencing) independently during posttest.	Well-designed AI scaffolding can yield durable cognitive benefits.
Experimental group consistently outperformed control group.	Theme 5: Shifting Perceptions of AI's Role in Professional Practice	Teachers viewed AI as a thinking partner rather than a shortcut, reinforcing deeper engagement and better outcomes.	Positive professional stances enhance effective and sustainable AI integration.

in the internalization of AI-supported strategies despite general positive outcomes.

Also, in the experimental group, the observed performance differences were driven by their engagement with the AI planning process. They used iterative prompting and compared AI-generated options, which aligns with their larger quantitative gains in creativity and lesson-plan quality. These cognitively demanding actions explained the substantial time × group interaction found in the MANOVA analysis. Control group teachers, using conventional methods, relied on familiar templates and routine decisions, limiting reflection and exposure to variations. Therefore, the AI-supported planning environment's richer cognitive engagement was central to the superior learning gains in the experimental participants.

Moreover, the comparative patterns observed in the posttest provide additional evidence of learning transfer. Although neither group had access to ChatGPT during the final assessment, the experimental group continued to outperform the control group on all three measures, suggesting that key elements of the AI-supported planning routines had been internalized. Qualitative findings confirmed that many experimental participants drew on strategies practiced during the intervention- such as systematically evaluating alternatives, articulating instructional rationales, and organizing activities around clearer pedagogical goals- despite the absence of AI assistance. This sustained advantage indicates that the intervention did not merely enhance short-term performance but facilitated deeper cognitive restructuring aligned with the principles of the transfer-of-learning framework. In contrast, control-group teachers reported approaching the posttest task similarly to the pretest, with limited evidence of new planning strategies or expanded reasoning processes. Taken together, the converging data show that guided interaction with AI can produce lasting cognitive and pedagogical effects that persist even when the technological scaffold is removed.

Consequently, the ChatGPT intervention significantly impacted teachers' cognitive and pedagogical performance. Quantitative improvements in critical thinking, creativity, and lesson-plan quality reflect deeper changes in how teachers made instructional decisions. Qualitative themes, such as increased cognitive engagement and a greater appreciation for AI's utility, explain these improvements.

Together, this suggests that structured AI interaction can enhance pedagogical reasoning, making it more reflective, flexible, and evidence-informed. Integrating such technological aids into teacher education is valuable, but guided practice is crucial for internalizing these benefits for long-term transfer.

The integrated evidence presented in Table 4 offers a coherent summary of how the quantitative outcomes and qualitative themes converge to explain the impact of the ChatGPT-assisted intervention. By mapping statistical gains onto the cognitive and pedagogical processes described by participants, the joint display clarifies the mechanisms through which the intervention influenced professional reasoning and instructional design. This integrated synthesis provides the conceptual foundation for the broader theoretical and practical interpretations elaborated in the following Discussion section.

INTERPRETATION OF RQ3 FINDINGS

Teachers' reflections revealed a nuanced set of perceived benefits and limitations regarding the role of ChatGPT in lesson planning. Many participants in the experimental group reported notable improvements in planning efficiency, explaining that the tool enabled them to generate initial instructional ideas more rapidly and allocate more time to refining pedagogical decisions. Several teachers also emphasized gains in instructional originality, describing how ChatGPT exposed them to alternative activity types, scaffolding techniques, and sequencing options that they had not previously considered. At the same time, interviewees from both the experimental and control groups expressed concerns about professional autonomy. While teachers valued the support provided by ChatGPT, they cautioned that excessive reliance on AI-generated suggestions could diminish their confidence in making independent instructional judgments. These insights highlight a complex interplay between the tool's capacity to enhance creativity and streamline planning and the ongoing need for educators to critically evaluate AI-generated content to avoid overdependence.

The mixed-methods synthesis reveals that the ChatGPT intervention enhanced teachers' planning through both direct and indirect mechanisms. Quantitatively, the significant time × group interaction, corroborated by MANOVA and follow-up ANOVA results, indicates that experimental teachers developed lasting cognitive routines rather than relying on temporary support. Qualitatively, these improvements are explained by increased metacognitive monitoring, iterative evaluation of alternatives, and deliberate decision-making. Ultimately, the AI served as a catalyst for deeper cognitive engagement, which helps explain why the experimental group consistently outperformed the control group.

Taken together, these quantitative, qualitative, and integrated findings provide a comprehensive account of how the ChatGPT-assisted intervention fostered teachers' cognitive and pedagogical development. The evidence demonstrates that improvements such as enhanced reasoning, creative exploration, and internalized planning were driven by specific cognitive processes that persisted beyond direct AI support. These findings establish a solid empirical foundation for understanding how AI-mediated scaffolding influences professional thinking. Consequently, the following discussion evaluates how these results align with or challenge existing literature while exploring their broader theoretical and practical significance for teacher education and AI-supported instructional design.

DISCUSSION

The results of this study contribute to a rapidly evolving body of research on AI-supported instructional planning. Consistent with findings by Karakaş and Yeşilyurt (2023) and Zhai et al. (2024), the current study demonstrates that ChatGPT can broaden teachers' pedagogical repertoires by generating diverse instructional pathways and facilitating flexible thinking. The strong improvement in creativity corresponds with Zhu et al. (2024), who reported that generative AI encourages open-ended exploration and enables teachers to experiment more readily with novel task designs. The qualitative evidence from the present study further reinforces the idea that structured exposure to AI-generated alternatives enhances teachers' capacity for imaginative adaptation.

This enhanced adaptability, fostered by AI's ability to generate multiple variations of instructional sequences, cultivates a more dynamic and responsive approach to teaching—one aligned with cognitive apprenticeship models that emphasize iterative refinement and practical problem-solving. Furthermore, exploring AI-generated options encourages teachers to consider a wider array of pedagogical moves, moving beyond conventional approaches and potentially reframing their understanding of effective instructional design. This process not only supports innovation but also addresses practical challenges by providing readily adaptable solutions, thereby integrating AI as a valuable tool for both creative expansion and pragmatic instructional adjustments.

Regarding critical thinking, the findings partially align with Liu et al. (2023) and Yıldız (2024), who observed that AI-supported environments can stimulate analytic engagement when teachers actively interrogate AI suggestions. However, the present study extends these insights by demonstrating that such gains are not automatic; rather, they depend on how teachers interact with the tool. Teachers who compared options, questioned assumptions, and engaged in reflective refinement showed substantial improvements, whereas more passive users displayed only modest gains. This divergence underscores the critical role of instructional design in AI-enhanced professional development. Explicit pedagogical strategies that foster structured inquiry and critical evaluation of AI outputs are crucial for maximizing benefits, as suggested by Guo and Lee (2023) and De la Puente et al. (2024).

These scaffolding interactions effectively transform the AI tool from a mere information provider into a cognitive mirror, compelling users to articulate their pedagogical

reasoning in order to justify or reject AI-generated outputs. Consequently, the cognitive workload shifts from rote planning to high-level evaluation, where the teacher's intentionality becomes the primary driver of quality. In this study, thoughtful and guided interaction- rather than unstructured use- appears to be the key factor that prevents overreliance and promotes cognitive development. Ultimately, these findings underscore that AI-integrated professional development must prioritize the cultivation of AI-literate skepticism as a foundational skill for sustaining long-term professional autonomy.

The observed transfer of learning is particularly significant. Few studies have examined whether AI-supported reasoning persists once the tool is removed or whether improvements extend beyond the immediate task environment. The current findings suggest that repeated cycles of comparison, justification, and refinement can help teachers internalize specific reasoning routines that remain functional outside the AI context. Importantly, this persistence appears contingent on how teachers engage with the tool: when interaction is structured around critique, explanation of choices, and revision- rather than passive acceptance- teachers are more likely to convert AI suggestions into personally meaningful cognitive strategies. In this way, the AI environment functions as a scaffold for learning-by-justification, corresponding with cognitive apprenticeship perspectives while also addressing a gap that empirical work has largely overlooked in teacher planning. Accordingly, the present study extends prior evidence by providing measurable and sustained effects on both planning quality and the quality of teachers' reasoning, offering a more durable mechanism for understanding how AI-supported professional learning translates into long-term practice.

Building upon the observed transfer of learning, the integration of AI in lesson planning holds the potential to foster sustained professional growth and cultivate adaptive expertise among educators. As teachers repeatedly engage in cycles of critique, justification, and refinement with AI tools, they not only improve immediate lesson designs but also develop metacognitive strategies for self-directed professional learning. This process encourages a shift from simply acquiring new techniques to developing a deeper understanding of pedagogical principles, enabling teachers to navigate unforeseen challenges and diverse student needs with greater flexibility and innovation long after the AI tool is no longer in direct use.

Looking ahead, the insights gained from this study suggest that AI-powered professional development is not merely about enhancing current practices but also about preparing teachers for the future of education. As educational landscapes become increasingly dynamic- with shifts in student demographics, advances in technology and evolving pedagogical theories- the skills developed through AI-literate skepticism and critical engagement, such as nuanced evaluation, structured inquiry and metacognitive awareness, will become indispensable.By equipping teachers with these capabilities, AI can serve as a catalyst for developing resilient and forward-thinking educators capable of independently adapting and innovating within complex teaching environments.

CONCLUSION AND IMPLICATIONS

This study investigated the impact of ChatGPT-assisted lesson planning on EFL teachers' critical thinking, creativity, and overall lesson plan quality. Both quantitative and qualitative data revealed substantial cognitive and pedagogical benefits from structured AI engagement, with the experimental group demonstrating more advanced planning competencies than those using traditional methods. Significant improvements in creativity and notable gains in critical thinking suggest that AI integration not only streamlines planning but also refines teachers' reasoning processes. Qualitative findings indicate that ChatGPT catalyzed reflective decision-making, prompting teachers to articulate rationales, compare alternatives, and justify choices more precisely. These evaluative cycles appear to have fostered the internalization of analytic routines, as evidenced by persistent posttest improvements even without direct AI assistance. Conversely, control group teachers reported routine-based planning, aligning with minimal changes in their quantitative outcomes.

Creativity showed the most pronounced and consistent improvement. Interview data suggested that teachers utilized ChatGPT as a prompt for idea divergence, adaptation, and expansion, rather than a template for replication. This explains the larger effect size for creativity compared to critical thinking, as AI-generated variations readily stimulate creative ideation. Critical thinking, however, requires deeper interrogation of assumptions and consequences, behaviors that varied among experimental teachers based on their active engagement with AI suggestions. Thus, the mixed-methods evidence suggests AI-supported planning uniformly strengthens creativity, while critical thinking gains are contingent on teachers' reflective stance during AI interaction.

A significant finding relates to teachers' evolving perceptions of AI. Initially viewing ChatGPT as a peripheral shortcut, sustained guided engagement led most experimental teachers to reframe it as a "thinking partner" that supports, rather than replaces, pedagogical reasoning. This shift in professional perception is crucial, as teachers' beliefs about technology significantly influence the realization of its affordances. Such changes, coupled with measurable cognitive gains, indicate that structured AI integration can foster more informed and balanced technology use in professional contexts.

The implications for teacher education are substantial. First, the findings underscore the importance of guided, scaffolded AI integration over incidental use. Second, teacher-training programs should embed reflective prompting strategies to encourage critical evaluation of AI-generated content. Finally, the evolving teacher perceptions highlight the need for professional development addressing both the affordances and risks of generative tools, promoting a balanced stance that maximizes benefits while

preserving professional autonomy.

Despite its contributions, the study's modest sample size, limited to in-service EFL teachers within a specific context, may constrain generalizability. The four-week intervention, while comparatively extended, might be insufficient to capture long-term cognitive development patterns. Furthermore, while the mixed-methods design offered rich insights, the qualitative phase relied on self-reported perceptions, potentially influenced by recall bias or participants' views of the research. Future research could benefit from longitudinal designs tracking the persistence of internalized reasoning strategies across different tasks. Comparative studies examining various AI scaffolding types could clarify which supports foster deep thinking most effectively. Investigating engagement across different teacher expertise levels would also be valuable. Finally, further exploration into the ethical, authorship, and authenticity dimensions of extensive AI use in instructional design is warranted.

Authors' contributions

All authors have contributed equally to prepare the paper.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflict of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Akrami, F., & Ghaderi, M. (2024). Transformation in teachers' professional knowledge with the emergence of artificial intelligence. Journal of Research in Curriculum Studies, 4(6), 43-64. https://doi.org/ 10.48310/JCDR.2024.16814.1109

Association of American Colleges and Universities. (2009). VALUE Rubrics. https://www.aacu.org/value-rubrics

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser ,453–494. Lawrence Erlbaum Associates.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and Conducting Mixed Methods Research (3rd ed.). SAGE Publications.

Cukurova, M., Kralj, L., Hertz, B. & Saltidou, E. (2024). Professional development for teachers in the age of AI. European Schoolnet. Brussels, Belgium.

Dahri, N. A., Yahaya, N., & Al-Rahmi, W. M. (2024). Exploring the influence of chatgpt on student academic success and career readiness. Education and Information Technologies. https://doi.org/10.1007/s10639-024-13148-2

De La Puente, M., Torres, J., Troncoso, A. L. B., Meza, Y. Y. H., & Carrascal, J. X. M. (2024). Investigating the use of chatgpt as a tool for enhancing critical thinkin and argumentation skills in international relations debates among undergraduate students. Smart Learning Environments, 11(25). https://doi.org/10.1186/s40561-024-00347-0

Den Berg, G. V., & Du Plessis, E. (2023). Chatgpt and generative AI: possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. Educ. Sci., 13(10), 998. https://doi.org/10.3390/educsci13100998

Do Amaral, I. (2024). Reflection on the use of generative language models as a tool for teaching design. IEEE World Engineering Education Conference(EDUNINE), Kos, Greece, 1-4. DOI:10.1109/EDUNINE60625.2024.10500634

Guo, Y., & Lee, D. (2023). Leveraging chatgpt for enhancing critical thinking skills. Journal of Chemical Education, 100,4876-4883. https://doi.org/10.1021/acs.jchemed.3c00505

Karakaş, A., & Yesilyurt, Y. E. (2023). The use of chatgpt for lesson planning. ChatGPT in Foreign Language Education and Translation Studies, 111-130. https://www.researchgate.net/publication/382063354

Kartal, G. (2024). The influence of ChatGPT on thinking skills and creativity of EFL student teachers: a narrative inquiry. Journal of Education for Teaching International research and pedagogy, 50 (4), 627-642.

Li, X., & Ni, X. (2024). Exploring the potential of ChatGPT for creative pedagogy: A mixed-methods study on teacher cognition and instructional design. Computers & Education, 208, 105889. https://doi.org/10.1016/j.compedu.2024.105889

Liu, Z. Y., Vobolevich, A., & Oparin, A. (2023). The influence of AI chatgpt on improving teachers' creative thinking. International Journal of Learning, Teaching and Educational Research, 22(12), 124-139. https://doi.org/10.26803/ijlter.22.12.7

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276-282.

Ng, D. T. K., Lee, M., Tan, R. J. Y., Hu, X., Downie, J. S., & Chu, S. K. W. (2023). A review of AI teaching and learning from 2000 to 2020. Education and Information Technologies, 28(7), 8445-8501. https://doi.org/10.1007/s10639-022-11491-w

Paul, R., & Elder, L. (2019). Critical Thinking: Tools for Taking Charge of Your Professional and Personal Life (3rd ed.). Pearson.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In International Encyclopedia of Education (2nd ed.). Pergamon Press

Pişkin Tunç, M. (2024). Examining pre-service mathematics teachers' purposes of using chatgpt in lesson plan development. Sakarya University Journal of Education, 14 (2), 391- 406. https://doi.org/10.19126/suje.1476326

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

Richards, J. C. (2013). Creativity in language teaching. Iranian Journal of Language Teaching Research, 1(3), 19-38.

Sawyer, R. K. (2011). Explaining Creativity: The Science of Human Innovation (2nd ed.). Oxford University Press.

Stognieva, O. (2024). Using chatgpt in planning English language classes. Informatics and Education,39(4):77-89. https://doi.org/10.32517/0234-0453-2024-39-4-77-89

Toma, R. B., & Yánez-Pérez, I. (2024). Effects of chatgpt use on undergraduate students' creativity: a threat to creative thinking?. Discover Artificial Intelligence, 4(74). https://doi.org/10.1007/s44163-024-00172-x

Wu, Y. (2024) Revolutionizing learning and teaching: crafting personalized, culturally responsive curriculum in the AI era. Creative Education, 15, 1642-1651. doi: 10.4236/ce.2024.158098.

Yildiz, T. A. (2024). Exploring the impact of ChatGPT on improving 21st-century skills for future English teachers during lesson planning. Computers in The Schools. DOI:10.1080/07380569.2024.2429534

Zhai, X., Romero, M., & Frøsig, T. B. (2024). Teacher agency in the age of generative AI: Towards a hybrid-intelligence framework for learning design. arXiv. https://doi.org/10.48550/arXiv.2407.06655

Zhua, S., Wangb, Z., Zhuangc, Y., Jiangd, Y., Guoe, M., Zhangf, X., & Gaod, Z. (2024). Exploring the impact of chatgpt on art creation and collaboration: benefits, challenges and ethical implications. Telematics and Informatics Reports, 14. https://doi.org/10.1016/j.teler.2024.100138