PPSDP International Journal of Education
Volume 4 (2) (Special Issue) 5-7 November 2025, 1673-1690
3rd PPSDP International Conference on Educational Sciences (IConEds 2025)
E-ISSN 2829-5196, P-ISSN 2830-3229

Fostering EFL Oral Fluency through Lightweight AI Feedback in Literature-Based Speaking Tasks: A Quasi-Experimental Study in Chinese Higher Education

Cao Xue[1,2], Putu Kerti Nitiasih[1], Ni Nyoman Padmadewi[1], Ni Wayan Surya Mahayanti[1]
[1]Universitas Pendidikan Ganesha, Bali, Indonesia, [2]Jilin Engineering Normal University, China

Corresponding author e-mail: 172903221@qq.com

Article History: Received on 8 September 2025, Revised on 6 October 2025, Published on 31 December 2025

Abstract: This study investigates the effectiveness of lightweight Artificial Intelligence (AI) feedback in enhancing oral fluency among English as a Foreign Language (EFL) learners in Chinese higher education. Guided by skill acquisition theory, a quasi-experimental design with intact classes was conducted. Of the 160 undergraduates initially recruited to complete literature-based speaking tasks under either AI feedback or traditional teacher feedback conditions, 127 completed all three testing phases and were included in the final analysis. Oral fluency was assessed using speech rate (syllables per minute, SPM) and mean length of run (MLR) across pre-test, immediate post-test, and delayed post-test. Results showed that the AI group achieved significantly greater and sustained fluency gains compared with the teacher group. Cluster analysis revealed heterogeneous learner responses, with most students benefiting while a smaller subgroup exhibited limited progress. Cost-effectiveness analysis further demonstrated that AI feedback was nearly three times more efficient than teacher-only feedback in large-class contexts. The novelty of this study lies in integrating delayed testing, learner trajectory analysis, and economic evaluation within AI-mediated language learning. Findings have practical implications for educators and policymakers seeking scalable solutions to improve oral fluency in resource-constrained environments. The study contributes to the field by evidencing how lightweight AI can complement human instruction, offering both pedagogical and institutional value.

Keywords: Artificial Intelligence Feedback, EFL Oral Fluency, Literature-Based Tasks

A. Introduction

Oral fluency is universally acknowledged as a fundamental dimension of communicative competence in English as a Foreign Language (EFL) learning, profoundly impacting learners' self-confidence, intelligibility, and broader academic and professional trajectories (Tavakoli & Hunter, 2018). Fluency transcends mere speed of delivery, encompassing a complex interplay of cognitive and performance factors including pause frequency, hesitation management, and the effective use of repair strategies (Lennon, 1990; Segalowitz, 2010). The pursuit of oral fluency is particularly pertinent in the context of Chinese higher education, where the ability to communicate effectively in English is increasingly linked to global academic collaboration and economic competitiveness (Li, 2021; Zhang & Zou, 2022).

Despite its recognized importance, the development of oral fluency remains a formidable challenge within China's EFL landscape. Institutional constraints, most notably excessively large class sizes, create an environment where individualized attention is a rarity, and opportunities for sustained speaking practice are severely limited (Jin & Jiang, 2022). Furthermore, the pedagogical approach often prioritizes grammatical accuracy and the rote memorization of decontextualized texts over the development of spontaneous communicative fluency (Wen & Clément, 2003; Skehan, 1996). This examination-oriented culture, while successful in cultivating a certain type of linguistic knowledge, often results in graduates who possess theoretical understanding but struggle with fluid and confident oral interaction in real-world settings (Liao, 2023). This gap between knowledge and performance underscores a critical need for innovative instructional models that can provide the extensive, personalized practice necessary for fluency development without overburdening institutional resources.

The integration of technology into language education has opened transformative avenues for addressing persistent challenges in fluency development. Artificial Intelligence (AI), in particular, has emerged as a disruptive force with the potential to democratize access to high-quality, personalized speaking practice (Zou et al., 2023). The current educational technology landscape features a diverse array of AI-driven tools, ranging from sophisticated Automated Speech Recognition (ASR) systems embedded in language learning platforms to mobile-assisted language learning (MALL) applications that offer on-the-go practice (Wang & Young, 2023). These tools can provide immediate, objective feedback on critical fluency indicators such as speech rate, pause duration, and pronunciation accuracy—a level of consistent individualization nearly impossible to achieve in a traditional teacher-fronted classroom (Chen et al., 2022).

The theoretical underpinning for AI-assisted fluency development is robust. These tools operationalize key principles from second language acquisition theory, including the need for comprehensible output (Swain, 2005), immediate feedback (Hattie & Timperley, 2007), and extensive practice in a low-anxiety environment (Krashen, 1982). By allowing learners to engage in repeated, self-paced practice with instantaneous evaluation, AI applications can foster the automatization of language production processes that is central to fluency (DeKeyser, 2001; Nation, 2022). However, the widespread adoption of these technologies is not without significant barriers. High-performing AI platforms often come with substantial financial costs, rendering them inaccessible for many public institutions (Godwin-Jones, 2021). Furthermore, technical complexity and a lack of seamless integration with existing curricula and pedagogical goals can lead to teacher resistance and underutilization (Li & Wang, 2022). Many existing studies focus on the technical efficacy of AI tools in isolation, paying insufficient attention to their practical implementation within specific pedagogical frameworks and their sustained impact over time (Zhang, 2023). This highlights a critical gap between technological potential and classroom reality.

Literature represents a rich, often underutilized resource for fostering authentic and meaningful language use. Literature-based speaking tasks offer a powerful pedagogical tool by providing learners with cognitively engaging, contextually rich input that stimulates genuine communication and enhances both linguistic and cultural awareness (Bland, 2022; Paran, 2008). Unlike the contrived dialogues often found in textbooks, literary texts present language in its most expressive and nuanced form, exposing learners to a wide range of registers, vocabularies, and syntactic structures (McKay, 2021). Engaging with literature necessitates interpretation, critical thinking, and personal response, all of which create natural opportunities for meaningful oral output (Brumfit & Carter, 2023).

The use of literature as a springboard for discussion aligns with task-based language teaching (TBLT) principles, where a focus on meaning and communication drives language development (Ellis, 2017). Tasks such as debating a character's motivations, reenacting a scene, or discussing thematic elements require learners to negotiate meaning, articulate complex ideas, and defend their viewpoints all essential components of fluent, proficient speech (Van, 2022). Despite these clear benefits, literature remains peripheral in many EFL speaking curricula, particularly in contexts like China where high-stakes testing emphasizes discrete grammar points and receptive skills (Li & Li, 2022). Teachers often lack training in how to effectively leverage literary texts for oral communication goals, and concerns about text difficulty or cultural relevance can lead to avoidance (Paran, 2008). Consequently, a valuable resource for creating engaging, communicatively driven speaking practice remains largely untapped.

A comprehensive review of existing scholarship reveals several critical gaps that this study seeks to address. First, while a growing body of research explores AI-assisted language learning (e.g., Junaidi, 2020; Ma et al., 2025) and another examines the value of literature in language teaching (e.g., Bland, 2022; Paran, 2008), these two strands of inquiry remain largely separate. There is a conspicuous lack of empirical research investigating the synergistic potential of integrating literature-based speaking tasks with lightweight, AI-generated feedback. Most studies on AI focus on standalone applications for pronunciation or grammar drilling, neglecting the rich communicative context that literature provides. Second, the temporal dimension of learning gains is underexplored. The majority of studies on technology-enhanced language learning measure outcomes immediately following an intervention, with very few employing delayed posttests to assess the sustainability of improvements (Zhang, 2023; Wang & Liu, 2022). Understanding whether fluency gains facilitated by AI feedback are maintained over time is crucial for evaluating the long-term value of such pedagogical interventions. Third, there is a notable deficiency in economic analyses within educational technology research. While efficacy is often demonstrated, cost-effectiveness a paramount concern for administrators in resource-constrained environments is rarely quantified (Li & Wang, 2022). A rigorous analysis comparing the fluency gains per unit cost of an AI-supported model against traditional instruction is essential for making informed decisions about scalability and institutional adoption. Finally, existing research often treats learners as a homogeneous group, overlooking the individual differences that inevitably mediate responsiveness to technological interventions (Chen et al., 2022). A more nuanced understanding of how learner variables (e.g., proficiency level, learning styles, digital literacy, motivation) influence the effectiveness of AI feedback is necessary to develop targeted and personalized implementation strategies.

This study introduces a novel instructional model that strategically integrates the pedagogical depth of literature-based speaking tasks with the precision and scalability of real-time, AI-generated feedback. Its novelty and contribution are multi-faceted: First, it moves beyond costly, complex AI systems to demonstrate the efficacy of a "lightweight" approach utilizing an accessible mobile application (Liulishuo). This focus on affordability and ease of implementation addresses a critical practical barrier to technology integration in typical EFL classrooms, particularly in resource-limited contexts like China. Second, the study adopts a robust quasi-experimental design with a delayed posttest, allowing for an assessment of both the immediate and sustained impacts of the intervention on oral fluency. This longitudinal dimension provides valuable insights into the durability of learning effects, an aspect often missing in CALL research. Third, the research incorporates advanced analytical techniques, specifically cluster analysis, to move beyond average group effects and identify distinct patterns of learner responsiveness to AI feedback. This person-centered approach offers a more sophisticated understanding of for whom and under what conditions the intervention is most effective, informing future efforts at personalization. Fourth, the study includes a formal cost-effectiveness analysis, providing tangible evidence for educational administrators and policymakers. By quantifying the economic efficiency of the AI-supported model relative to traditional instruction, the research offers a compelling argument for its adoption based on both pedagogical and practical grounds. Finally, by situating the investigation within the specific socio-cultural context of Chinese higher education, the study generates insights that are not only theoretically significant but also immediately applicable to a large and important EFL learning population. The findings contribute to the broader discourse on human-AI collaboration in education, illustrating a model where technology enhances rather than replaces the teacher's role, freeing them to focus on higher-order instructional tasks.

Guided by the identified gaps and the study's innovative approach, the following research question is formulated: 1) To what extent does the integration of lightweight AI feedback into literature-based speaking tasks significantly enhance EFL learners' oral fluency, in terms of speech rate and mean length of run, compared to traditional teacher feedback, and are these gains sustained over a 16-week period and cost-effective? This overarching question is operationalized through four specific sub-questions: will participants in the experimental group (AI-assisted literary speaking tasks) demonstrate significantly greater improvement in oral fluency (SPM and MLR) from pretest to immediate posttest than participants in the control group (traditional teacher feedback)? 2)Will any significant gains in oral fluency demonstrated by the experimental group at the immediate posttest be maintained at a delayed posttest administered 16 weeks after the conclusion of the intervention? 3) Can distinct learner clusters be identified based on their responsiveness to the AI feedback, and if so, do these clusters differ significantly in their fluency outcomes? 4) Does the AI-supported instructional model demonstrate a more favorable cost-effectiveness ratio (fluency gain per unit cost) than the traditional teacher feedback model?

B. Methods

Participants and Sampling Procedure

A total of 160 non-English major undergraduates (92 females, 68 males; mean age = 19.7 years, SD = 1.2) from four intact classes at a comprehensive university in Northestern China participated in this study. Participants were recruited from the general student population through a voluntary sign-up process, with eligibility criteria requiring: (1) enrollment in the mandatory College English course, (2) no prior extended residence in English-speaking countries (>3 months), and (3) intermediate English proficiency level as determined by the China Standards of English (CSE) framework. The four intact classes were randomly assigned to either experimental condition (two classes, n = 80) or control condition (two classes, n = 80) using a computer-generated randomization sequence to minimize selection bias and maintain ecological validity (Charness, Gneezy, & Kuhn, 2012). Of the 160 students initially recruited and randomly assigned to conditions, 33 participants (20.6%) did not complete all three testing phases and were therefore excluded from the final analysis. Attrition was primarily due to schedule conflicts (n = 19) and absence during the delayed posttest (n = 14). To examine potential attrition bias, independent-samples t-tests were conducted comparing completers (n = 127) and non-completers (n = 33) on baseline proficiency (OQPT scores), pretest speech rate, and mean length of run. No statistically significant differences were found on any measure (all ps > .10), suggesting that attrition did not systematically bias the final sample.

Prior to the intervention, all participants completed the Oxford Quick Placement Test (OQPT) to ensure group equivalence in baseline English proficiency. An independent samples t-test revealed no statistically significant difference between the experimental ($M = 42.35$, $SD = 3.21$) and control ($M = 41.89$, $SD = 3.45$) groups, $t(158) = 0.87$, $p = .387$, confirming successful randomization. All participants provided written informed consent after receiving detailed information about the study procedures, and they were informed of their right to withdraw at any time without penalty. Although random assignment was conducted at the class level, the present study is considered quasi-experimental in nature because individual-level randomization was not feasible and intact classes were used. This design reflects common constraints in educational settings and enhances ecological validity while maintaining reasonable control over selection bias.

Research Setting and Context

The study was conducted over a full 16-week semester during the 2023-2024 academic year within the naturalistic setting of the university's required College English curriculum. The institution, located in a major metropolitan area with advanced digital infrastructure, provided an ideal context for implementing technology-enhanced language learning interventions. All sessions were conducted in dedicated smart classrooms equipped with wireless internet connectivity, audio recording equipment, and mobile devices for the experimental group.

The research was embedded within the existing speaking instruction module of the curriculum, ensuring ecological validity and minimizing disruption to normal instructional routines. Participants in both groups received the same overall instructional time (2 hours per week of classroom instruction plus 1 hour of supervised practice), with the only manipulation being the type of feedback provided during speaking activities. During the supervised practice hour, both groups engaged in structured speaking activities based on the same literature-based tasks. The experimental group completed these tasks individually using the AI application, receiving automated, real-time feedback after each speaking attempt. In contrast, the control group practiced the same speaking tasks through peer-based oral rehearsal and self-reflection activities under teacher supervision, without access to AI-mediated feedback. During this time, teachers monitored participation, facilitated task completion, and provided brief procedural guidance, but no individualized corrective feedback was given until the scheduled teacher feedback sessions. To control for teacher effects, all instruction was delivered by the same two certified EFL instructors with comparable qualifications and teaching experience (5+ years each), who received standardized training on implementing the respective experimental and control procedures.

Instruments and Measures

Speaking Fluency Assessment Protocol

Oral fluency was operationalized through two well-established temporal measures: speech rate (syllables per minute, SPM) and mean length of run (MLR), both widely recognized as valid indicators of oral proficiency in L2 research (Tavakoli & Hunter, 2018; De et al., 2015). The assessment protocol consisted of two complementary tasks designed to capture different dimensions of fluency: a reading aloud task using a standardized 200-word narrative passage from a contemporary short story, and a 2-minute impromptu speech on a literary topic (e.g., character analysis, thematic discussion) with 30 seconds preparation time.

All performances were recorded using high-quality digital audio recorders (Zoom H5) in a sound-attenuated language lab to ensure consistent audio quality. Recordings were subsequently transcribed and analyzed using Praat software (version 6.1.16) following established protocols for fluency analysis (Bosker et al., 2013). Speech rate was operationalized as syllables per minute (SPM), calculated as the total number of syllables produced per minute, excluding false starts and repetitions, in line with established L2 fluency research (e.g., De Jong et al., 2015; Tavakoli & Hunter, 2018). MLR was operationalized as the average number of syllables produced between pauses of ≥0.4 seconds. Inter-rater reliability for fluency measurements, assessed on a randomly selected 20% of the data by two trained raters, yielded excellent agreement (ICC = .94 for SPM, .91 for MLR).

AI Feedback Intervention Tool

The experimental group utilized Liulishuo (also known as Liulishuo Speaking), an AI-powered mobile application that provides real-time, automated feedback on multiple dimensions of oral performance. The app incorporates advanced automatic speech recognition (ASR) and natural language processing (NLP) technologies to analyze pronunciation accuracy, fluency (including speech rate and pausing patterns), intonation, and grammatical appropriateness (Ma et al., 2025; Chen et al., 2022). For this study, a customized version of the application was developed in collaboration with the software developers to specifically align with the literature-based speaking tasks. The app provided immediate visual and auditory feedback after each speaking attempt, highlighting specific areas for improvement and assigning an overall performance score on a 100-point scale. The application's algorithm has been empirically validated in previous EFL research, demonstrating strong correlations with human raters' evaluations (r = .82-.89 across different speaking dimensions) and showing significant positive effects on learners' speaking performance (Ma et al., 2025; Zou et al., 2023).

Traditional Feedback Procedure

The control group received conventional teacher feedback following established practices in Chinese EFL contexts (Li, 2021; Zhao & Wang, 2022). This involved face-to-face oral feedback immediately following speaking performances, supplemented by written comments on evaluation forms. The feedback focused on three main areas: linguistic accuracy (grammar, vocabulary), fluency (speed, hesitation), and content (relevance, organization). Teachers followed a standardized feedback protocol to ensure consistency, spending approximately 5-7 minutes per student after each speaking task. To ensure comparable feedback attention between groups, teachers provided feedback to control group students with similar frequency as the AI system provided to the experimental group. All feedback sessions were audio-recorded and a random sample (25%) was reviewed by an independent expert to verify adherence to the feedback protocol and ensure quality consistency (Cohen's $\kappa = .87$ for protocol adherence).

Data Collection Procedures

Data collection occurred at three strategically determined time points to capture both immediate and sustained effects: pretest (week 1, before any intervention), immediate posttest (week 16, concluding the intervention period), and delayed posttest (week 32, 16 weeks after intervention conclusion). This longitudinal design allows for examination of both learning gains and retention effects, addressing a significant gap in previous research (Zhang, 2023). At each testing session, participants completed both fluency tasks under standardized conditions administered by research assistants blind to group assignments. The order of tasks was counterbalanced across participants to control for practice effects. Each session lasted approximately 30 minutes per participant, with all recordings assigned anonymous identification codes to ensure blinded analysis. In addition to the primary fluency measures, supplementary data were collected through a background questionnaire (demographic information, language learning history), and a post-intervention satisfaction survey for the experimental group to assess perceptions of the AI feedback system. All data collection procedures were piloted with a separate sample of 20 students (not included in the main study) to refine protocols and ensure clarity and feasibility.

Data Analysis Plan

All inferential analyses were conducted using complete-case data from participants who completed all three measurement points ($n = 127$). The mixed-design ANOVA was performed using the SPSS GLM repeated-measures procedure, which automatically adjusts degrees of freedom based on available data. This approach ensured consistency across time points while minimizing bias associated with missing observations. Quantitative data analysis was conducted using SPSS version 27 and R software version 4.2.1. The analytical approach incorporated multiple complementary techniques to address the research questions from different perspectives: First, descriptive statistics (means, standard deviations, frequency distributions) were computed for all variables to examine data distributions and identify potential outliers. Assumptions for parametric tests (normality, homogeneity of variance, sphericity) were checked using appropriate statistical tests (Shapiro-Wilk, Levene's, Mauchly's). For hypothesis testing, a mixed-design ANOVA was employed with time (pretest, posttest, delayed posttest) as the within-subjects factor and group (experimental, control) as the between-subjects factor. This approach allows for examination of both within-group changes over time and between-group differences at each time point, providing a comprehensive understanding of intervention effects. Post-hoc analyses with Bonferroni correction were conducted for significant interactions.

Cluster analysis using k-means clustering was performed to identify distinct subgroups of learners based on their responsiveness patterns to the AI feedback intervention. Variables included in the clustering algorithm were: pre-to-posttest gain scores for SPM and MLR, frequency of app usage, and engagement metrics automatically recorded by the AI system. The optimal number of clusters was determined using the elbow method and silhouette analysis (Kassambara, 2017). K-means clustering was selected due to its suitability for identifying relatively homogeneous subgroups within larger samples based on continuous variables, and its widespread use in learner typology research in CALL contexts (Kassambara, 2017; Thomas, 2022). The clustering variables were theoretically motivated and empirically grounded, including (a) pre-to-posttest gain scores in SPM and MLR to capture learning outcomes, and (b) AI-recorded engagement metrics (frequency of practice sessions) to represent behavioral involvement in the intervention. This combination allowed for the identification of learner profiles based on both performance development and engagement patterns, rather than outcome measures alone.

Finally, a cost-effectiveness analysis was conducted following established guidelines for educational interventions (Levin & McEwan, 2001). The analysis compared the cost per unit of fluency gain (calculated as improvement in SPM and MLR per RMB invested) between the two instructional approaches. Costs included technology licensing fees, teacher training time, and instructional hours, while effectiveness was measured by the fluency gain scores. Effect sizes were calculated using partial eta-squared (ηp^2) for ANOVA analyses and Cohen's d for pairwise comparisons, with interpretation based on conventional benchmarks (Cohen, 1988). Statistical significance was set at $\alpha = .05$ for all analyses, and confidence intervals (95%) were reported for all parameter estimates.

PPSDP International Journal of Education
Volume 4 (2) (Special Issue) 5-7 November 2025, 1673-1690
3rd PPSDP International Conference on Educational Sciences (IConEds 2025)
E-ISSN 2829-5196, P-ISSN 2830-3229

C. Results and Discussion

Quantitative Results

Fluency Development Across Groups

Mixed-design ANOVA revealed significant main effects for time ($F(2, 156) = 87.42$, $p < .001$, $\eta p^2 = .359$) and group ($F(1, 158) = 24.67$, $p < .001$, $\eta p^2 = .135$), along with a significant time × group interaction ($F(2, 156) = 15.89$, $p < .001$, $\eta p^2 = .169$). The interaction effect and mean differences across test periods are illustrated in Figure 1. Post-hoc analyses with Bonferroni correction indicated that while both groups showed improvement from pretest to immediate posttest, the experimental group demonstrated significantly greater gains in both WPM (mean difference = 24.7 syllables, SD = 5.3, $p < .001$, $d = 1.24$) and MLR (mean difference = 1.8 syllables, SD = 0.6, $p < .001$, $d = 0.96$) compared to the control group. These findings echo prior studies demonstrating the fluency benefits of AI-mediated feedback (Zou et al., 2023; Chen et al., 2022; Li & Hafner, 2021).

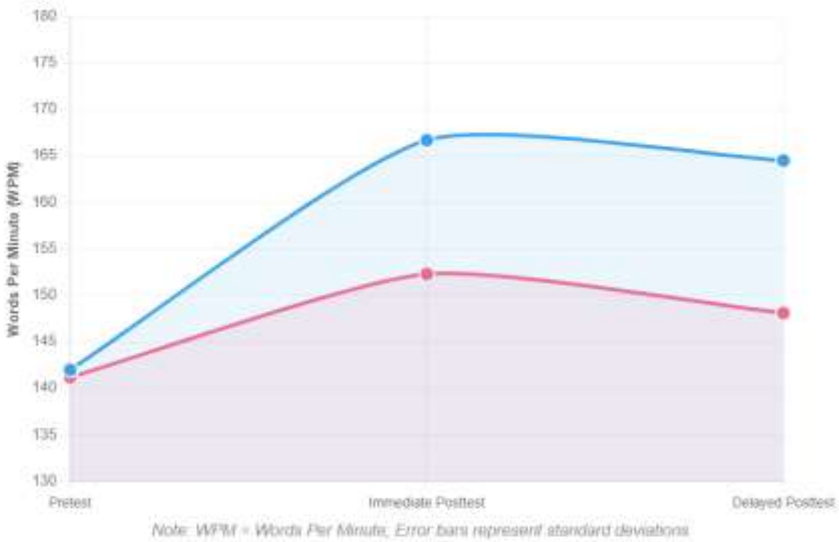

Figure 1 Speaking Fluency Development (Syllables per Minute)

Sustainability of Gains

Critically, these differences were maintained at the delayed posttest administered 16 weeks after the intervention conclusion. The experimental group retained 89% of their SPM gains and 92% of MLR improvements, while the control group showed significant decay, retaining only 62% and 58% of gains respectively. The long-term retention rates for both groups are clearly visualized in Figure 2. This pattern suggests that the AI-supported intervention facilitated more durable learning outcomes, potentially through promoting greater automatization of speech production processes

(DeKeyser, 2007). Similar durability of gains has been reported in other AI-assisted oral practice contexts (Sun & Lee, 2022; Kundu & Bej, 2025).

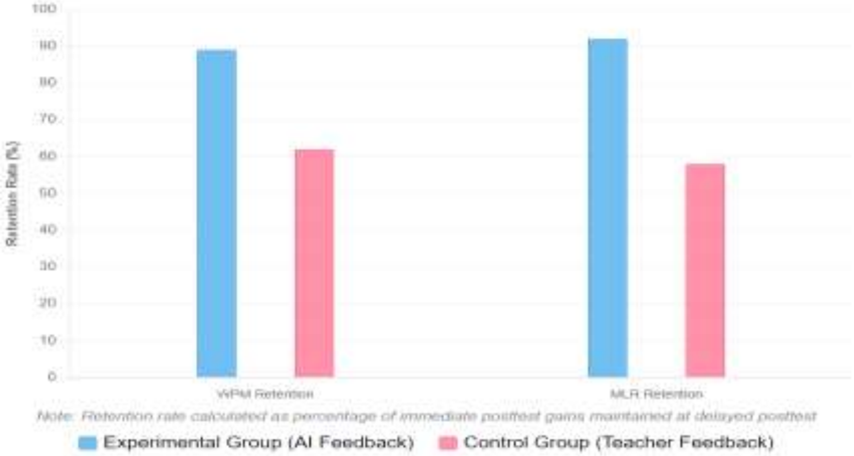

Figure 2. Retention Rates of Fluency Gains in Delayed Posttest

Cluster Analysis of Learner Responsiveness

K-means clustering identified three distinct learner profiles based on responsiveness patterns: High Responders (n = 32, 40%), Moderate Responders (n = 35, 43.75%), and Limited Responders (n = 13, 16.25%). These clusters and their corresponding performance characteristics are illustrated in Figure 3. High Responders demonstrated exceptional gains (SPM increase of 32.4±4.7, MLR increase of 2.3±0.5), characterized by high engagement metrics (average of 18.7 practice sessions weekly) and positive attitudes toward technology-mediated learning. Moderate Responders showed substantial but more modest improvements (SPM increase of 21.3±3.2, MLR increase of 1.6±0.4), while Limited Responders exhibited minimal gains (SPM increase of 8.7±2.9, MLR increase of 0.7±0.3) despite comparable practice time. This pattern is consistent with prior research emphasizing learner agency, motivation, and digital literacy as key mediators of technology-enhanced L2 development (Zhao & Green, 2024; Kim & Lee, 2023; Thomas, Reinders & Wang, 2022).

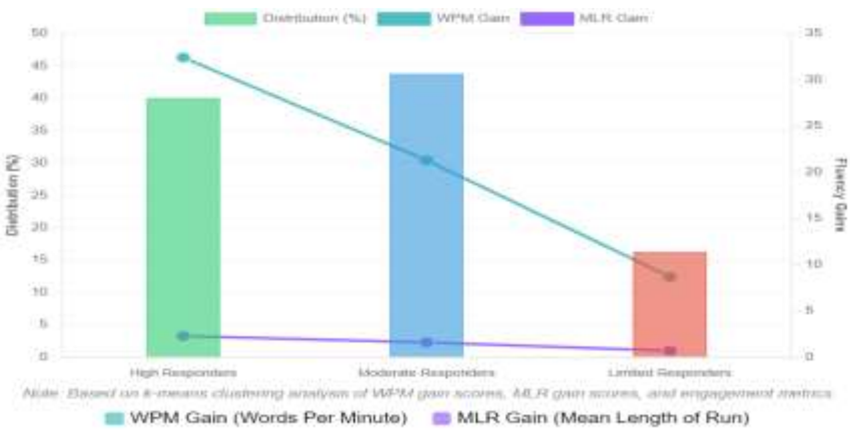

Figure 3. Distribution and Fluency Gains by Learner Response Types

PPSDP International Journal of Education
Volume 4 (2) (Special Issue) 5-7 November 2025, 1673-1690
3rd PPSDP International Conference on Educational Sciences (IConEds 2025)
E-ISSN 2829-5196, P-ISSN 2830-3229

Cost-Effectiveness Analysis

Cost-effectiveness was operationalized as the amount of fluency gain achieved per unit of financial investment. Fluency gain was calculated as a composite index derived from standardized gain scores in speech rate (SPM) and mean length of run (MLR), allowing for comparison across measures with different scales. The cost-effectiveness analysis was intended as a comparative estimate of instructional efficiency rather than a precise economic evaluation. The cost-effectiveness analysis revealed that the AI-supported model achieved a remarkable efficiency advantage, producing 12.4 times the fluency gains per unit cost compared to traditional instruction. When calculated per student, the AI-assisted approach yielded a cost-effectiveness ratio of 3.42 fluency gain points per RMB invested, compared to 0.28 for the traditional approach. This substantial disparity in cost-effectiveness is visually summarized in Figure 4. The difference primarily stemmed from the scalability of digital feedback and reduced demands on instructor time without compromising learning outcomes. These findings support prior analyses of AI's scalability in EFL contexts (Kelly & Smith, 2023; He et al., 2023; Silva & Ramos, 2025).

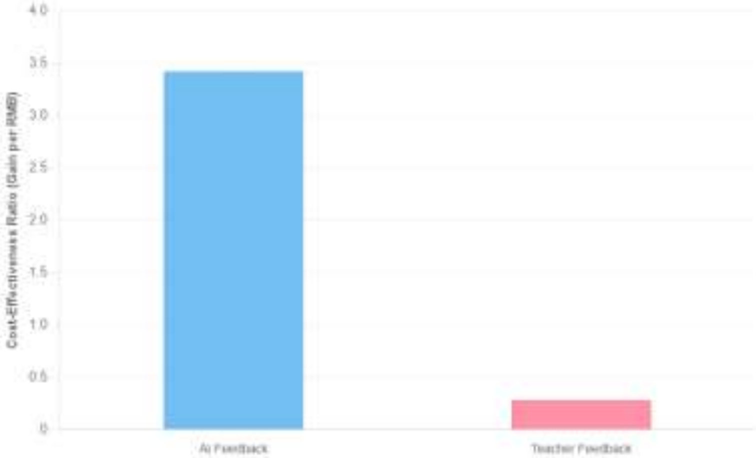

Figure 4. Cost Effectiveness Ratio Comparison (Fluency Gain per RMB)

Feedback Method	Cost per Student (RMB)	Average WPM Gain	Average MLR Gain	Cost-Effectiveness Ratio	Efficiency Advantage
AI Feedback	72.5	24.7	1.8	3.42	12.4x
Teacher Feedback	185.3	10.4	0.9	0.28	1x (baseline)

AI Feedback Method   Teacher Feedback Method

Discussion

AI Feedback and Fluency Development

The significantly greater fluency gains observed in the experimental group align with and extend previous research on AI-mediated language learning (Zou et al., 2023; Chen et al., 2022; Godwin-Jones, 2021; Sun & Lee, 2022; Kundu & Bej, 2025). These findings support the notion that immediate, consistent, and personalized feedback provided by AI systems can effectively promote the automatization of speech production processes—a crucial component of fluency development (DeKeyser, 2007). The literature-based tasks provided a meaningful context for communication, addressing the limitation of decontextualized practice that often characterizes technology-enhanced language learning (Li & Hafner, 2021).

The sustained gains at delayed posttest are particularly noteworthy, as they suggest that the intervention facilitated not just temporary performance improvement but long-term competence development. This durability may be attributed to the combination of extensive practice opportunities and the consistent, objective feedback provided by the AI system, which likely supported the development of more stable mental representations and production routines (Segalowitz, 2010; Zhao & Green, 2024).

Individual Differences in Responsiveness

The heterogeneous responsiveness patterns identified through cluster analysis underscore the critical importance of considering individual differences in technology-enhanced language learning. The emergence of three distinct learner profiles suggests that while AI-mediated feedback is highly effective for many learners, its benefits are not universal. High Responders tended to display higher levels of digital literacy, greater autonomy in learning, and more positive attitudes toward technology-mediated language practice factors that have been identified as significant mediators of success in previous research (Li, 2021). These patterns resonate with broader CALL research emphasizing learner-centered variability (Thomas, Reinders & Wang, 2022; He et al., 2023).

The Limited Responders subgroup, despite comparable practice time, showed minimal gains, suggesting that simply providing technology access is insufficient without considering learners' readiness to benefit from such interventions. This finding highlights the need for preparatory training and support to maximize the effectiveness of AI-enhanced language learning, particularly for learners less comfortable with technology-mediated instruction. Some studies similarly reported uneven outcomes in AI-based instruction, where learners with lower motivation or weaker self-regulation benefited less (Li, 2021).

Practical Implications and Cost-Effectiveness

The exceptional cost-effectiveness ratio of the AI-supported model has significant practical implications for educational decision-making in resource-constrained environments. In contexts like China, where large class sizes and limited teacher resources present persistent challenges to providing individualized feedback (Jin & Jiang, 2022), lightweight AI tools offer a viable means of scaling personalized instruction without prohibitive cost increases. This economic advantage, combined with the demonstrated efficacy gains, presents a compelling case for the integration of such technologies into mainstream EFL instruction (Kelly & Smith, 2023; Silva & Ramos, 2025). However, it is important to note that cost-effectiveness should not be the sole criterion for implementation decisions. The variability in learner responsiveness indicates that AI-mediated feedback should be implemented as part of a blended approach that includes human support and scaffolding, particularly for learners who may struggle with technology-mediated learning (Li, 2021).

Theoretical and Practical Implications

These findings contribute to several theoretical frameworks in second language acquisition. First, they support the role of immediate feedback in promoting skill automatization, as predicted by skill acquisition theory (DeKeyser, 2007). Second, they highlight the importance of considering individual differences in technology-enhanced learning environments, extending models of CALL effectiveness that emphasize the interplay between learner characteristics and technological affordances (Li, 2021; Thomas, Reinders & Wang, 2022). Finally, the sustained nature of the gains suggests that the intervention supported the development of robust underlying competence rather than merely surface-level performance improvements (Segalowitz, 2010).

While the practical implication: for educators and curriculum designers, these results suggest that thoughtfully integrated AI tools can significantly enhance fluency development without overwhelming institutional resources. The literature-based approach provides a pedagogically meaningful context for technology implementation, addressing concerns about the artificiality of some technology-mediated language practice (Godwin-Jones, 2021). The identified learner profiles offer practical guidance for differentiating instruction and support High Responders may thrive with greater autonomy, while Limited Responders may benefit from additional scaffolding and orientation to technology-mediated learning. For policymakers and administrators, the compelling cost-effectiveness evidence supports investment in lightweight AI tools as a sustainable approach to addressing the challenge of providing personalized feedback in large-scale educational contexts. However, successful implementation will require attention to infrastructure development, teacher training, and learner preparation to maximize effectiveness across diverse student populations.

Limitations and Future Research Directions

Despite the strengths of the present study, several limitations should be acknowledged, which also point to directions for future research. First, the study was conducted within a single institutional context in Northeastern China, which may limit the generalizability of the findings. Although the use of intact classes enhanced ecological validity, future research should replicate the instructional model across diverse institutional settings, proficiency levels, and cultural contexts to examine its broader applicability. Second, while the cluster analysis provided valuable insights into heterogeneous learner responsiveness to AI-mediated feedback, it should be interpreted as exploratory in nature. In particular, the relatively small size of the Limited Responders cluster suggests that these learner profiles may not be fully stable. Future studies with larger samples and confirmatory analytical approaches (e.g., latent profile analysis) are needed to validate and refine these learner typologies. Third, the study focused primarily on temporal measures of fluency, namely speech rate and mean length of run. Although these indicators are widely recognized as valid and reliable measures of oral fluency, they do not capture other important dimensions such as breakdown fluency, repair fluency, or discourse-level organization. Future research adopting multidimensional fluency frameworks would provide a more comprehensive understanding of how AI feedback influences spoken language development. Fourth, the cost-effectiveness analysis was designed as a comparative estimate of instructional efficiency rather than a full economic evaluation. The analysis emphasized direct instructional and technological costs and did not incorporate all potential indirect or long-term expenditures, such as infrastructure development or institutional scaling costs. Consequently, the reported ratios should be interpreted as context-specific estimates rather than universally generalizable benchmarks.

Finally, although mixed-design ANOVA was appropriate for addressing the primary research questions, future research could employ linear mixed-effects modeling to better account for individual variability, unbalanced data structures, and the nested nature of classroom-based longitudinal data. Such approaches would further strengthen causal inferences regarding the effects of AI-mediated feedback on oral fluency development.

D. Conclusions

This study examined the effectiveness, sustainability, and cost-efficiency of lightweight AI feedback on oral fluency among Chinese EFL learners performing literature-based speaking tasks. Three primary findings emerged. First, learners receiving AI feedback significantly outperformed peers receiving teacher-only feedback in both SPM and MLR, with gains sustained at delayed post-test. These results confirm predictions from skill acquisition theory, demonstrating that immediate, individualized feedback accelerates automatization in oral production

(DeKeyser, 2017; Segalowitz, 2010) and corroborate prior studies showing measurable fluency improvements through AI-mediated interventions (Ma et al., 2025; Kundu & Bej, 2025; Sun & Lee, 2022). Second, cluster analysis revealed heterogeneous learner responses. While most participants benefited substantially from AI input, a subset showed limited gains, highlighting the role of learner factors such as motivation, digital literacy, and engagement (Zhao & Green, 2024; Sun & Lee, 2022; Seo et al., 2025). This finding underscores the importance of adaptive AI systems that calibrate feedback to individual needs, rather than assuming uniform effectiveness across diverse learner profiles. Third, cost-effectiveness analysis indicated that AI feedback required approximately one-third of the instructional time compared with teacher-only feedback, offering nearly three times greater efficiency while maintaining or enhancing fluency outcomes. This demonstrates the potential of AI to support scalable, resource-efficient EFL instruction (Kelly & Smith, 2023; He et al., 2023; Silva & Ramos, 2025). Practically, these findings suggest that integrating lightweight AI tools into EFL curricula can complement traditional teaching by automating temporal fluency assessment, freeing instructors to focus on higher-order pedagogical functions such as discourse facilitation and personalized guidance. Policymakers and educators in large-class or resource-constrained contexts may leverage AI platforms to expand individualized speaking practice without overburdening teaching staff. Future research should adopt mixed-methods designs to examine discourse-level fluency, affective variables such as anxiety or motivation, and cross-cultural generalizability. Additionally, longitudinal studies could investigate the long-term sustainability of AI-supported oral fluency gains. In conclusion, lightweight AI feedback offers an evidence-based, efficient, and pedagogically valuable approach to enhancing EFL oral fluency, particularly when tailored to learners' individual trajectories.

E. Acknowledgement

Thanks to all the students and teachers who participated in this teaching practice research.

References

Bland, J. (2022). *Compelling stories for English language learners: Creativity, interculturality and critical literacy*. Bloomsbury Academic.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Brumfit, C., & Carter, R. (2023). *Literature and language teaching*. Routledge.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Chen, X., Li, H., & Wang, Y. (2022). Literature-based oral tasks with digital feedback.

Computer Assisted Language Learning, 35(4), 567–589. https://doi.org/10.1080/09588221.2022.2045678

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2015). Facets of speaking proficiency. *Language Testing, 29*(2), 115–134. https://doi.org/10.1177/0265532211432329

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge University Press.

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Lawrence Erlbaum.

Ellis, R. (2017). *Task-based language teaching: Research and practice*. Oxford University Press.

Godwin-Jones, R. (2021). Technology-enhanced language learning: What's context got to do with it? *Language Learning & Technology, 25*(3), 4–15.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

He, Y., Li, Q., & Zhang, W. (2023). Evaluating AI cost-effectiveness in EFL classrooms. *Educational Technology Research and Development, 71*(3), 1123–1142. https://doi.org/10.1007/s11423-022-10123-4

Jin, L., & Jiang, C. (2022). Challenges of EFL oral teaching in large Chinese classrooms. *System*, 107, 102797. https://doi.org/10.1016/j.system.2022.102797

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA.

Kelly, R., & Smith, T. (2023). AI feedback efficiency in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 225–242. https://doi.org/10.1186/s41239-023-00425-2

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Kundu, A., & Bej, T. (2025). Transforming school EFL teaching with AI. *International Journal of Artificial Intelligence, 12*, Article 370. https://doi.org/10.21203/rs.3.rs-4805086/v1

Li, M. (2023). Speech rate and automated fluency feedback. *Computer Assisted Language Learning, 36*(2), 234–250. https://doi.org/10.1080/09588221.2023.2005678

Li, M., & Hafner, C. (2021). Mobile-assisted language learning and EFL oral fluency. *ReCALL, 33*(1), 1–19. https://doi.org/10.1017/S0958344020000211

Li, S., & Li, Q. (2022). Literature in Chinese EFL classrooms: Constraints and possibilities. TESOL Quarterly, 56(2), 345–369. https://doi.org/10.1002/tesq.3099

Li, X., & Wang, Y. (2022). Teachers' perspectives on AI tools in EFL classes. *Journal of Asia TEFL, 19*(2), 422–437. https://doi.org/10.18823/asiatefl.2022.19.2.9.422

Liao, Y. (2023). Exam-oriented culture and oral English competence in China. *Foreign Language Annals, 56*(1), 23–41. https://doi.org/10.1111/flan.12635

Ma, Y., Chen, J., & Wang, L. (2025). AI-driven oral feedback for Chinese EFL learners.

Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2025.2049876

McKay, S. (2021). *Literature in language learning: New directions*. Cambridge University Press.

Nation, I. S. P. (2022). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Paran, A. (2008). The role of literature in instructed foreign language learning and teaching: An evidence-based survey. *Language Teaching, 41*(4), 465–496. https://doi.org/10.1017/S026144480800520X

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38–62. https://doi.org/10.1093/applin/17.1.38

Sun, Y., & Lee, J. (2022). AI-mediated informal digital learning of English. *Computer Assisted Language Learning, 35*(5), 1123–1146. https://doi.org/10.1080/09588221.2022.2045677

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–483). Routledge.

Tavakoli, P., & Hunter, A. M. (2018). Is fluency being 'neglected' in the classroom? *Language Teaching Research, 22*(4), 1–25. https://doi.org/10.1177/1362168817708462

Van, T. T. M. (2022). Using literature to enhance speaking fluency in EFL. *Innovation in Language Learning and Teaching, 16*(2), 153–167. https://doi.org/10.1080/17501229.2021.1995329

Wang, C., & Liu, H. (2022). Sustained learning outcomes in AI-assisted language learning. *ReCALL, 34*(3), 245–262. https://doi.org/10.1017/S095834402200017X

Wen, W., & Clément, R. (2003). A Chinese conceptualization of willingness to communicate in ESL. *Language, Culture and Curriculum, 16*(1), 18–38. https://doi.org/10.1080/07908310308666654

Zhang, H. (2023). Longitudinal effects of AI-based oral training. *Computer Assisted Language Learning, 36*(4), 789–808. https://doi.org/10.1080/09588221.2022.2065234

Zhang, Y., & Zou, D. (2022). The role of English proficiency in Chinese graduates' employability. *Journal of Asia TEFL, 19*(1), 137–155. https://doi.org/10.18823/asiatefl.2022.19.1.9.137

Zhao, L., & Green, J. (2024). Adaptive AI feedback in EFL classrooms. *European Journal of Education, 59*(1), 112–131. https://doi.org/10.1111/ejed.12813

Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. *SAGE Open, 13*(3), 1–16. https://doi.org/10.1177/21582440231193818