Issues and Strategies in Teaching Listening and Speaking Skills within the Framework of Foreign Language Pedagogy

[1]Manzura Shamsitdinova Gopporovna, [2]Dildora A. Usarova, [3]Abdullayeva Dilfuza Muxiddinovna, [4]Bektosheva Durdonakhon Bakhtiyorovna, [5]Ibragimova Shokhsanam Rashid qizi, [6]Vakhobova Marguba Abdukhamidovna, [7]Eshchanova Gulbakhor, [8]Iskhakova Fotima Sharifovna,

[1]Associate professor, Foreign Languages Department of Tashkent State Law University, Tashkent (Uzbekistan). m.shamsitdinova@tsul.uz https://orcid.org/0000-0001-8078-3559

[2]PhD, Acting Associate Professor of English for Specific Purposes (ESP) Department of Foreign Languages, Tashkent State University of Law (TSUL)
Tashkent, Uzbekistan ORCID: https://orcid.org/0000-0001-9867-3673 Email: d.usarova@tsul.uz

[3]Acting Associate Professor of the Department of Foreign Languages PhD in Pedagogical Sciences, Tashkent State University of Law (TSUL) Email: abdullayeva.d@tsul.uz ORCID: https://orcid/org 0009-0008-8436-2338

[4]Acting Teacher of English for Specific Purposes (ESP) Department of Foreign languages, Tashkent State University of Law (TSUL) Tashkent, Uzbekistan Email: d.bektosheva@tsul.uz ORCID: https://orcid.org/0009-0007-7988-7765

[5]English teacher Department of Social and Exact Sciences Tashkent State University of Economics (TSUE) Email: shokhsanam9202@gmail.com

[6]Acting Senior lecturer of the Department of Social and Exact Sciences PhD in Philosophical Sciences Tashkent State University of Economics Email: vahobovama@gmail.com ORCID: https://orcid.org/0009-0006-7050-5110

, PhD [7]Head of the Department of Foreign Philology, Urgench State Pedagogical Institute, Uzbekistan ORCID: 0009-0009-2781-9730 E-mail: gulbahoreshchanova@gmail.com

[8]Assistant at the Department of Otolaryngology No. 1, Samarkand State Medical University, Samarqand, Uzbekistan E-mail: fatima.isxakova@mail.ru https://orcid.org/0009-0009-4864-4108

HOW TO CITE:

Manzura Shamsitdinova Gopporovna, Dildora A. Usarova, Abdullayeva Dilfuza Muxiddinovna, Bektosheva Durdonakhon Bakhtiyorovna, Ibragimova Shokhsanam Rashid qizi, Vakhobova Marguba Abdukhamidovna, Eshchanova Gulbakhor, Iskhakova Fotima Sharifovna (2026). Issues and Strategies in Teaching Listening and Speaking Skills within the Framework of Foreign Language Pedagogy. International Journal of Special Education, 41(12s), 1165-1185

ABSTRACT:

Background: Listening and speaking are mutually constitutive in communicative competence, yet classroom practice often treats them as discrete skills. Recent studies (2020–2025) underscore the value of dialogic tasks, multimodal input, metacognitive regulation, and AI-supported personalization for enhancing perceptual processing, interactional competence, and fluency within ecologically valid learning environments.

Research gap: Empirical work rarely integrates cognitive models of speech perception with interactional pragmatics and task-based pedagogy in a single design. Moreover, designs that coordinate adaptive input (e.g., calibrated speech rate, lexical density) with real-time conversational feedback remain under-specified, and evidence on transfer across genres and modalities is limited.

Aim: To examine persistent issues in teaching listening and speaking and to evaluate strategy-rich, technology-enhanced interventions that align cognitive processing with communicative, task-based principles.

Methodology: A mixed-methods, quasi-experimental study followed 168 university EFL learners over 12 weeks in four sections: CLT baseline; CLT+metacognitive prompting; CLT+AI-adaptive input and interactive chat; and a combined

condition. Data included standardized listening and speaking assessments (rater-moderated CEFR scales), clickstream traces (pauses, caption use), dialogic analytics (turn length, repair moves), and stimulated-recall interviews. Multilevel modeling estimated learning gains; thematic analysis examined strategy deployment.
Findings: The combined condition yielded the largest improvements in listening (g≈0.58) and speaking (g≈0.52). AI adaptivity reduced extraneous load and increased lexical coverage; metacognitive cycles (plan–monitor–evaluate) supported transfer of segmentation, inferencing, and interactional repair strategies. Learners demonstrated greater autonomy by calibrating caption density, pausing, and rehearsal timing; conversational feedback enhanced turn-taking fluidity and reduced dysfluency clusters.
Implications: Integrate adaptive difficulty, graduated captions, strategic pausing, and interactive feedback within task-based cycles, coupling explicit metacognitive scaffolds with opportunities for authentic interaction and reflective debrief.
Original contribution: The study articulates a unified, CLT-compatible framework that operationalizes cognitive and interactional strategy instruction with AI-mediated adaptivity, offering actionable design principles for scalable, genre-transferrable gains in L2 listening and speaking.

Keywords: listening comprehension, speaking fluency, communicative language teaching, metacognitive strategies, AI-assisted learning, task-based pedagogy, interactional competence, adaptive difficulty, captions, learner autonomy.

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors. Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Introduction

Oral communication remains the most consequential dimension of foreign language use in globalized education, mobility, and work. Whether negotiating meaning in transnational teams, participating in hybrid university seminars, or collaborating in digitally mediated projects, learners' capacity to listen and speak under real-time constraints determines their access to knowledge and social participation. The policy discourse across regions has accordingly elevated oracy—listening and speaking as coordinated competencies—into an explicit curricular priority, shifting assessment regimes and pedagogy toward interactional performance rather than discrete linguistic knowledge (Alexander, 2020; Mercer et al., 2022). Within applied linguistics, this evolution reflects a broader reorientation from decontextualized grammar teaching toward communicative, task-based, and interaction-driven approaches that foreground meaning-making, co-construction of discourse, and situated use of language (Ellis, 2020; Carless, 2022). Yet, despite decades of reform, classrooms still struggle to move beyond "speaking-as-display" and "listening-as-decoding" toward dialogic, strategy-rich instruction that enables learners to cope with authentic variability, noise, speed, and socio-pragmatic contingency (Field, 2021; Plonsky & Tavakoli, 2021).

Listening and speaking are not parallel skills but interdependent processes in the orchestration of communicative competence. Listening is prospective as much as retrospective: fluent interlocutors anticipate trajectories of talk, update situation models, and allocate attention to prosodic and pragmatic cues that shape turn-taking and repair (Cutler, 2021; Field, 2021). Speaking, conversely, is constrained by what has been heard—lexical activation is primed by input, and interactional moves are calibrated against

interlocutor uptake and stance (Kang & Johnson, 2021). Contemporary accounts of oracy conceptualize this coupling in terms of adaptive control: listeners predict and segment in order to respond; speakers monitor and modulate delivery in response to feedback, both verbal and multimodal (Mercer et al., 2022). This interdependence has methodological consequences. Pedagogies that isolate listening comprehension from dialogic production may improve item-level scores while failing to cultivate the timing, strategic pausing, inferencing under noise, and repair practices that sustain real conversation (Saito, 2021; Tavakoli & Wright, 2020). The present study takes this coupling seriously, arguing for integrated designs that treat comprehension and production as mutually enabling within communicative tasks.

The historical shift away from grammar-centered instruction (GCI) toward communicative language teaching (CLT) and task-based language teaching (TBLT) is well documented; however, the tension between structural coverage and communicative performance persists in policy and practice (Long, 2021; Ellis, 2020). GCI prioritized accuracy, sequence, and form-referenced syllabi; CLT/TBLT reoriented goals toward meaning-focused interaction, functional outcomes, and learner agency. Yet "hybrid drift" is common: classrooms adopt communicative tasks but retain accuracy-first epistemologies that discourage risk-taking and hinder strategic listening and interactional fluency (Carless, 2022). Moreover, standardized testing regimes can inadvertently reify monologic speaking and decontextualized listening, resulting in pedagogical washback that sidelines interactive skills (Harding & Brunfaut, 2020). Recent intervention research suggests that explicit strategy instruction embedded within authentic tasks can mediate this tension, enabling learners to mobilize bottom-up and top-down processes while engaging in purposeful interaction (Goh & Vandergrift, 2022; Cross, 2023). What is less clear is how to combine metacognitive scaffolds with adaptive technologies and dialogic tasks to deliver scalable gains in both listening and speaking across genres and modalities.

Learners face a constellation of challenges in developing oral communication. Linguistically, dense lexical neighborhoods, unfamiliar collocations, reduced forms, and variable accents elevate processing demands (Saito, 2021; Boers, 2021). Cognitively, segmentation at natural speech rates competes with limited working memory; listeners must reconcile rapid phonetic variation with probabilistic lexical prediction under time pressure (Cutler, 2021). Psychologically, anxiety and low self-efficacy distort attention and trigger maladaptive avoidance strategies, undermining both comprehension and fluency (Jiang & Dewaele, 2020; Teimouri, Goetze, & Plonsky, 2020). Technologically, digital environments can either alleviate or exacerbate load: captions and playback controls support processing yet risk over-reliance if poorly faded; automated feedback can guide repair but is uneven in sensitivity to pragmatics (Roquet et al., 2021; Li & Hegelheimer, 2021). Socioculturally, participation norms, face concerns, and power dynamics shape turn-taking and willingness to communicate, especially in lingua franca and intercultural settings where pragmatic routines diverge (Kecskes, 2022; Fang, 2022). The compounded effect is that many learners plateau: they can answer comprehension questions and deliver rehearsed monologues but falter in spontaneous interaction where rapid inferencing, flexible formulation, and micro-repair are needed.

Technological ecosystems have changed the opportunity structure for oracy development. Mobile-assisted language learning (MALL) now affords micro-learning, ubiquitous access to authentic audio, and push-based practice integrated into daily routines (Burston, 2020; Viberg & Grönlund, 2021). Synchronous and asynchronous digital communication environments—video conferencing, voice messaging, and social audio—sustain authentic, multimodal interaction, though they also introduce latency, overlapping speech, and platform-specific conventions that learners must master (Hampel, 2021; Jin, 2022). Meanwhile, AI-supported language learning (AILL) has progressed beyond drill to adaptive delivery and

conversational agents that can calibrate difficulty (speech rate, lexical density), generate targeted prompts, and provide immediate feedback on intelligibility, prosody, and turn construction (Cukurova et al., 2022; Gao, Wang, & Jiao, 2023). Automated speech recognition (ASR) and large language model (LLM)-driven chat interfaces enable low-stakes, anytime speaking practice and reflective listening tasks that externalize comprehension processes (Li & Hegelheimer, 2021; Huang & Chen, 2024). The promise is compelling: scalable personalization, abundant input, and immediate, data-informed feedback. Yet research cautions that naive adoption can misalign with communicative goals, privileging form-focused or single-turn metrics (accuracy, word count) at the expense of interactional competence, repair, and pragmatic appropriateness (Park & Son, 2022; Lai & Zheng, 2023).

Theoretical development has kept pace. Contemporary cognitive models conceive listening as active inference, where listeners integrate acoustic detail, lexical probabilities, and contextual schemas to update a situation model in real time (Cutler, 2021; Field, 2021). This aligns with metacognitive perspectives emphasizing planning, monitoring, and evaluation cycles that enable learners to select and adapt strategies—segmenting, shadowing, tolerance of ambiguity—under authentic variability (Goh & Vandergrift, 2022). Interactional competence frameworks, in turn, foreground co-construction: the ability to manage turn-taking, enact repair, and deploy stance and alignment resources appropriate to context (Hall, Hellermann, & Pekarek Doehler, 2021; Saito, 2021). A pedagogical synthesis suggests that classrooms should stage tasks that create conditions for prediction and segmentation, scaffold metacognitive self-regulation, and cultivate interactional moves within meaningful activity. The challenge is operational: how to design tasks, supports, and assessment that instantiate this synthesis in ways that scale.

Empirically, several trends since 2020 invite cautious optimism. Studies report that graduated captioning—adjusting density and timing and then fading support—improves comprehension and transfer better than all-or-nothing caption use (Roquet et al., 2021; Montero Perez, 2022). Metacognitive instruction embedded in listening cycles enhances strategic awareness and self-efficacy while yielding measurable gains on complex input (Goh & Vandergrift, 2022; Cross, 2023). In speaking, interventions that train prosodic fluency and repair practices lead to improvements in comprehensibility and interactional smoothness rather than merely increasing speed (Saito, 2021; Tavakoli & Wright, 2020). MALL deployments show that spaced micro-tasks and reflective prompts can sustain engagement and broaden exposure to accent diversity (Viberg & Grönlund, 2021; Burston, 2020). AILL studies suggest that ASR-based feedback on segmentals/suprasegmentals and LLM-based dialogic practice can increase time-on-task and reduce anxiety, though effects vary by alignment with task goals and the granularity of feedback (Cukurova et al., 2022; Park & Son, 2022; Huang & Chen, 2024). What is missing is integrative evidence showing that combined designs—metacognitive scaffolds plus AI adaptivity within dialogic tasks—produce durable, genre-transferrable gains in both listening and speaking.

Several gaps in the literature remain salient. First, the integration gap: many studies target either listening or speaking in isolation, leaving the oracy interface under-theorized and under-tested in classroom settings (Plonsky & Tavakoli, 2021). Second, the design specificity gap: while calls for metacognitive scaffolding are common, concrete design principles for sequencing supports (e.g., when to introduce, calibrate, and fade captions; how to time strategic pausing; how to engineer interactive prompts that externalize comprehension) are unevenly articulated (Cross, 2023; Goh & Vandergrift, 2022). Third, the adaptivity gap: research on AI-mediated personalization often measures proximal outcomes (word error rate, accuracy on immediate tasks) without tracing effects on interactional competence, repair, and transfer to novel genres or noisy conditions (Lai & Zheng, 2023; Gao et al., 2023). Fourth, the equity and

ethics gap: studies rarely examine how algorithmic feedback interacts with accent diversity, multilingual identities, and participation rights, risking technological washback that privileges standard varieties and constrains learner voice (Kecskes, 2022; Mercer et al., 2022). Finally, the assessment gap: formative analytics and conversational metrics are emerging, but validated measures that capture oracy as co-constructed performance remain limited in mainstream practice (Harding & Brunfaut, 2020; Hall et al., 2021).

Against this backdrop, the present article contributes a unified framework for teaching listening and speaking that couples cognitive strategy training and metacognitive regulation with AI-supported adaptivity and task-based, interactional pedagogy. Conceptually, we synthesize cognitive listening as active inference, metacognitive cycles of self-regulation, and interactional competence into a design space where supports are purposeful, transparent, and fade over time. Methodologically, we operationalize this synthesis in a quasi-experimental, mixed-methods study that contrasts baseline CLT with conditions featuring metacognitive prompting, AI-adaptive input and interactive feedback, and a combined design. We track outcomes not only on standardized listening and speaking measures but also on process indicators—caption density control, pausing and rehearsal timing, turn-taking fluency, and repair moves—and triangulate with learner accounts of strategy use and perceived autonomy.

The significance of this study is threefold. First, it advances theory by modeling oracy as a coupled system: listening and speaking are trained together through tasks engineered for prediction, segmentation, formulation, and repair, with supports calibrated to cognitive load and interactional demand (Field, 2021; Hall et al., 2021). Second, it advances design knowledge by specifying actionable principles for adaptive difficulty (speech rate, lexical density), captioning (timing, density, fading), strategic pausing and rehearsal, and conversational feedback that foregrounds interactional competence rather than surface accuracy (Montero Perez, 2022; Saito, 2021). Third, it advances equity by problematizing accent and genre coverage, encouraging exposure to diverse speaker identities, and interrogating how AI-mediated feedback can be tuned to avoid normative bias while still providing usable guidance (Kecskes, 2022; Mercer et al., 2022). In a landscape saturated with tools, we prioritize principled pedagogy: technology serves communicative ends and cognitive/interactional development, not the reverse.

Research Objectives and Questions

Guided by the foregoing synthesis, the study pursues the following objectives:

1. To identify persistent issues that constrain the development of listening and speaking in mainstream foreign language classrooms, including linguistic, cognitive, psychological, technological, and sociocultural factors.

2. To design and evaluate strategy-rich, technology-enhanced interventions that integrate metacognitive scaffolding, adaptive input/feedback, and dialogic tasks to support coupled gains in listening and speaking.

3. To examine how learners regulate supports (e.g., caption density, pausing) and deploy interactional strategies (e.g., repair, turn construction) across genres, modalities, and levels of input variability.

4. To articulate design principles and assessment practices that are scalable, ethically attentive, and compatible with CLT/TBLT curricula.

From these objectives, we derive the following research questions:
RQ1: Compared with baseline CLT, to what extent do metacognitive scaffolding, AI-adaptive input/interactive feedback, and their combination differentially affect L2 learners' listening comprehension and speaking performance over a sustained instructional period?
RQ2: How do these interventions shape learners' process behaviors—strategic pausing, caption use and fading, rehearsal timing, turn-taking

fluency, and repair—and how do such behaviors mediate outcome gains?
RQ3: To what extent do gains transfer across genres (e.g., academic, transactional, casual) and modalities (audio-only, video, synchronous chat), and under what conditions (e.g., increased noise, unfamiliar accents) are effects sustained?
RQ4: How do learners experience and negotiate AI-mediated feedback with respect to autonomy, anxiety, and perceptions of fairness and accent inclusivity?

In addressing these questions, we aim to move the field beyond piecemeal innovations toward a coherent, empirically grounded account of how listening and speaking can be taught together, at scale, in ways that respect the complexity of language use and the diversity of learners. The study's originality lies in its coupling of theoretical depth with design specificity, its insistence on dialogic performance as the evaluative target, and its nuanced engagement with the promises and pitfalls of AI- and mobile-supported learning within real classroom ecologies (Ellis, 2020; Carless, 2022; Cukurova et al., 2022; Huang & Chen, 2024). By foregrounding orchestrated attention, strategic externalization of comprehension processes, and interactional responsiveness, the framework offered here seeks not merely to raise test scores but to cultivate durable, transferable oracy for participation in multilingual worlds.

Literature Review

Theoretical foundations of listening and speaking instruction

The theoretical underpinnings of listening and speaking instruction have broadened from structuralist and behaviorist antecedents toward cognitively and socially situated accounts of language use. Early form-focused paradigms treated listening as passive decoding and speaking as controlled production, prioritizing accuracy through decontextualized drills. Contemporary approaches integrate psycholinguistic processing, metacognitive regulation, and interactional perspectives, situating oracy within task-based, communicative ecologies (Ellis, 2020; Hall, Hellermann, & Pekarek Doehler, 2021). Three strands are especially consequential. First, cognitive models conceptualize listening as predictive, resource-limited processing in which segmentation, lexical access, and situation-model updating unfold under time pressure (Cutler, 2021; Field, 2021). Second, sociocultural and interactionist views frame speaking as co-constructed action—turn-taking, repair, stance, and alignment—shaped by participation norms and power asymmetries (Kecskes, 2022; Saito, 2021). Third, metacognitive theories foreground strategic self-regulation, arguing that planning, monitoring, and evaluation mediate transfer across tasks and genres (Goh & Vandergrift, 2022; Cross, 2023). The pedagogical implication is a synthesis that couples explicit strategy work with opportunities for authentic interaction and principled use of technological scaffolds—a stance that avoids a false dichotomy between meaning-focused use and form-focused support.

Oracy and communicative competence

The construct of oracy re-centers listening and speaking as mutually constitutive competencies enabling participation in academic, professional, and civic life (Alexander, 2020; Mercer, Warwick, & Ahmed, 2022). Communicative competence frameworks have diversified—from Canale and Swain's grammatical, sociolinguistic, and strategic components to interactional competence models that emphasize co-construction in real time (Hall et al., 2021). Recent scholarship positions oracy as distributed across individuals and tools: captions, transcripts, and conversational agents reshape how comprehension and production are coordinated (Mercer et al., 2022). However, tensions persist. Advocates of oracy argue for dialogic pedagogy and assessment reform, yet large-scale testing often privileges monologic tasks that misrepresent interactive performance (Harding & Brunfaut, 2020). Studies in school and tertiary contexts report gains from dialogic instruction on reasoning and language development (Mercer et al., 2022), but evidence specific to L2 oracy across multilingual settings is uneven, with variability in task authenticity and attention to pragmatic appropriateness. A critical synthesis suggests that oracy should be operationalized not as aggregated subskills but as

situated performance measured through task-sensitive, interactional indicators (e.g., repair initiation, timing of backchannels, turn relevance) and supported by explicit instruction in listening-for-action and speaking-for-alignment.

Listening comprehension models

Listening has shifted from a bottom-up versus top-down binary toward interactive, predictive accounts. Field (2021) reappraises classroom listening as managing decoding bottlenecks—phonological reduction, coarticulation—while allocating attention to discourse cues. Cutler (2021) emphasizes experience-dependent tuning of phoneme categories and lexical probabilities. Work on multimodal listening highlights how gesture, gaze, and on-screen text alter processing pathways (Montero Perez, 2022). Recent empirical studies (2020–2025) converge on three claims. First, segmentation and lexical retrieval are primary constraints at natural speech rates; interventions that slow, segment, or enhance salient cues yield near-term benefits but risk dependency if not faded (Roquet et al., 2021). Second, predictive processing is fragile under noise; training tolerance of ambiguity and inferencing from discourse structure supports resilience (Goh & Vandergrift, 2022). Third, multimodal input can scaffold or distract depending on timing and density; graduated captioning and strategically placed visuals enhance transfer compared to always-on supports (Montero Perez, 2022). Methodologically, many listening studies rely on comprehension questions that conflate decoding with higher-order inference. Process-tracing (eye-tracking, clickstream, shadowing) is more informative but underused; when employed, it reveals heterogeneous strategy profiles that correlate with outcomes (Cross, 2023). A persistent limitation is narrow genre coverage, with over-reliance on scripted materials that underrepresent accent diversity and spontaneous features central to real-world listening.

Speaking performance and interactional competence

Research on speaking has matured beyond speed and accuracy metrics to encompass comprehensibility, fluency dimensions (speed, breakdown, repair), and interactional resources (Saito, 2021; Tavakoli & Wright, 2020). Studies since 2020 underscore that prosodic fluency and effective repair predict listener judgments more than raw speech rate (Plonsky & Tavakoli, 2021; Saito, 2021). Interactional competence research highlights learners' abilities to manage turn-taking, sequence actions, and negotiate meaning through repair and alignment practices (Hall et al., 2021). Task design matters: dialogic tasks that require information exchange and perspective-taking elicit richer interactional moves than monologic prompts, but reliability in rating such performance is challenging (Harding & Brunfaut, 2020). Intervention studies report benefits from explicit instruction on formulaic sequences for openings/closings and stance-taking, coupled with reflective debriefs on repair episodes (Boers, 2021). Contradictions arise around the role of explicit phonetic training: some trials show gains in segmental accuracy without parallel improvements in interactional smoothness, suggesting that form work must be embedded within conversational practice to influence communicative outcomes (Saito, 2021). A methodological weakness is the sparse use of longitudinal designs; short interventions capture proximal gains while leaving durability and transfer under-examined.

Cognitive and metacognitive dimensions

Cognitive load theory and limited-capacity models explain why real-time oracy tasks overwhelm learners: parsing, lexical access, formulation, and monitoring compete for scarce resources. Metacognitive instruction can mitigate overload by externalizing processes and sequencing supports (Goh & Vandergrift, 2022). Recent randomized or quasi-experimental studies show that metacognitive cycles (plan–monitor–evaluate) embedded in listening and speaking tasks improve self-efficacy and performance, especially when learners receive feedback on strategy use (Cross, 2023). Evidence further suggests that strategy instruction yields the largest gains when tailored to task demands (e.g., segmenting and shadowing for dense audio; pre-emptive planning and turn-templates for fast

exchanges). However, the literature is not unequivocal: some studies report modest or non-significant gains, often where strategy instruction remains declarative, short in duration, or detached from assessment (Goh & Vandergrift, 2022). Measurement is another weakness: many studies infer metacognitive change from self-reports; fewer triangulate with process data (clickstream, think-aloud under time constraints) or delayed transfer tasks. A promising direction involves "scaffold fading" protocols that algorithmically reduce support as learners demonstrate control, thereby preventing over-reliance while preserving confidence.

Foreign language anxiety and affective variables

Affective factors—anxiety, enjoyment, grit, willingness to communicate—shape attention allocation, risk-taking, and persistence. Meta-analytic and large-sample studies link lower anxiety and higher enjoyment to better L2 performance, including speaking fluency and listening comprehension (Jiang & Dewaele, 2020; Teimouri, Goetze, & Plonsky, 2020). During and after pandemic-induced online learning, speaking anxiety often decreased for some learners due to lower social risk, while others reported increased cognitive load from managing platforms and self-monitoring (Jin, 2022). Interventions leveraging low-stakes, private practice with ASR or chat agents can reduce anxiety and increase willingness to communicate, but effects vary by perceived authenticity and feedback tone (Park & Son, 2022). Contradictions persist regarding captions: while they can reduce anxiety by buffering comprehension, they also induce dependence and self-blame when learners struggle without them, suggesting the need for graduated support and explicit affective framing. Methodologically, affect research often treats constructs as stable traits; newer work highlights situational variability and calls for micro-longitudinal designs to capture within-person dynamics under different task and technological conditions.

Technology-enhanced language learning

Technology-enhanced language learning (TELL) encompasses multimodal input, digital assessment, analytics, and social platforms. Since 2020, the literature documents rapid expansion in synchronous video tools, interactive video platforms, and learning analytics applied to oracy. Strengths include scalability, access to authentic content, and opportunities for multimodal scaffolding (Hampel, 2021). Yet the "pedagogy–technology gap" remains: tools proliferate faster than principled designs, leading to feature-led rather than goal-led adoption (Carless, 2022). Studies with stronger designs align technology with task-based cycles, integrate formative feedback, and report not only test outcomes but process and transfer measures (Roquet et al., 2021; Montero Perez, 2022). Weaknesses include brief interventions, convenience samples, non-equivalent controls, and reliance on self-report or immediate post-tests. The most robust evidence shows that pairing technology with explicit strategy work and scaffold fading produces durable benefits; technology alone yields mixed results. Ethical considerations—privacy, data bias, accessibility—are increasingly discussed but rarely tested empirically in classroom trials.

AI-assisted speaking and listening instruction

AI in oracy pedagogy spans ASR for feedback, text-to-speech with controllable parameters, and LLM-based conversational agents. ASR systems now provide granular feedback on intelligibility, segmental errors, and prosody, with some studies reporting improved comprehensibility and increased practice time (Li & Hegelheimer, 2021; Cukurova, Kent, & Luckin, 2022). More recent work explores adaptive difficulty—modulating speech rate, lexical density, and noise based on learner performance—and conversational analytics that prompt repair or elaboration in real time (Gao, Wang, & Jiao, 2023). LLM-driven agents can sustain topic continuity, simulate interlocutor variability, and offer metacognitive prompts, potentially externalizing listening and speaking processes (Huang & Chen, 2024). Evidence is promising but heterogeneous. Positive findings include increased time-on-task, reduced speaking anxiety, and gains in targeted features. Limitations include: (a) construct validity—systems often optimize for surface accuracy rather than interactional competence; (b) bias—feedback may privilege standard

accents and penalize legitimate varieties; (c) ecological validity—lab gains may attenuate in noisy, multiturn classroom talk; and (d) opacity—model decisions are hard to interpret pedagogically (Lai & Zheng, 2023). Stronger studies use mixed methods, triangulating automated metrics with human ratings of interactional quality and with learner reflections. A key research need is to evaluate AI as part of an orchestrated design, not as a standalone tutor.

Mobile-assisted language learning (MALL)

MALL research highlights ubiquitous access, micro-learning, and context-aware practice. Systematic reviews over two decades show steady positive effects on vocabulary and receptive skills, with growing attention to oracy (Burston, 2020; Viberg & Grönlund, 2021). Recent MALL studies for speaking integrate spaced prompts, voice messaging, and reflective journals; for listening, they deploy push-based micro-podcasts with embedded strategy cues. Methodological strengths include ecological validity and sustained engagement; weaknesses include self-selection bias and limited control conditions. Contradictions arise around autonomy: some learners flourish with self-paced micro-tasks, while others drift without social accountability. The most effective MALL designs embed social presence—peer response, teacher nudges—and metacognitive framing to connect micro-practice with classroom tasks. Accessibility and equity deserve emphasis: device heterogeneity, bandwidth constraints, and notification fatigue shape participation patterns, yet are underreported in efficacy claims.

Methodology

Research paradigm

This study is grounded in a pragmatic paradigm that privileges methodological appropriateness to the research questions and educational context over allegiance to any single epistemology. Pragmatism supports combining experimental control with ecological validity and values both quantifiable learning gains and interpretive accounts oflearner and teacher experience (Johnson & Onwuegbuzie, 2004; Morgan, 2014). Given the dual aims—estimating the effects of strategy-rich, technology-enhanced designs on listening and speaking, and understanding how learners regulate supports and enact interactional strategies—pragmatism provides a coherent rationale for mixed methods.

Research design

We employed a convergent, mixed-methods, quasi-experimental design with four parallel instructional conditions implemented over 12 weeks in intact university EFL classes. Conditions were: (C1) CLT baseline; (C2) CLT + metacognitive prompting; (C3) CLT + AI-adaptive input and interactive feedback; and (C4) combined supports (metacognitive prompting + AI adaptivity). Classes were randomly assigned to conditions at the section level to minimize contamination while preserving institutional timetables (Shadish, Cook, & Campbell, 2002). A pretest–posttest design captured change in standardized listening and speaking outcomes, while process data (clickstream traces; dialogic analytics) and qualitative evidence (observations, interviews) illuminated mechanisms. Quantitative and qualitative strands were given equal priority and integrated at interpretation through joint displays and process–outcome mediation models.

Mixed-methods justification

The causal-comparative component alone cannot explain how or why changes occurred; likewise, qualitative accounts without outcome measures risk limited generalizability. Mixed methods allow us to (a) estimate treatment effects; (b) examine process indicators (caption density control, pausing behavior, repair moves) as mediators; and (c) triangulate learner/teacher perspectives with trace and performance data, thereby enhancing internal validity, explanatory depth, and practical transferability (Creswell & Plano Clark, 2018). Convergence and divergence across strands were explicitly analyzed to test explanatory coherence.

Participants

Participants were 184 undergraduate EFL learners (first- and second-year non-English majors) at a public university. Inclusion criteria were: intermediate proficiency (B1–B2) as

determined by institutional placement tests; enrollment in required oral communication courses; consent to data collection. Exclusion criteria: prior participation in pilot studies; extended absence (>20% sessions). Demographics: 61% female; mean age 19.8 (SD=1.4). To mitigate cluster-level imbalances, we stratified by program (business, engineering, social sciences) before section-level assignment. Four instructors (two per faculty, 6–10 years' experience) delivered the interventions after common training. Sample size was powered a priori using Monte Carlo simulation for multilevel models (α=.05, power=.80) to detect small-to-moderate effects (f≈.20) accounting for intraclass correlation (ICC≈.05–.08).

Sampling techniques

We used cluster sampling at the class level with stratified randomization to conditions. For qualitative components, maximum-variation purposeful sampling selected 32 learners (8 per condition; balanced by gender and proficiency band) for interviews and targeted observations to capture a range of strategy profiles and technology dispositions (Palinkas et al., 2015). Teacher interviews included all four instructors. This sampling ensured representativeness in outcomes and depth in mechanisms.

Research setting

The study was conducted in blended courses meeting twice weekly (90 minutes) in technology-enabled classrooms and via the institution's learning platform. Each room had projection, reliable Wi-Fi, and laptops/tablets available upon request. The digital environment integrated: (a) an interactive video platform supporting graduated captions, variable playback, and embedded prompts; (b) an AI-enabled conversational practice tool providing adaptive speech rate/lexical density in listening tasks and feedback on prosody/comprehensibility in speaking; and (c) learning analytics to capture non-identifying clickstream and dialogic metrics. All tools complied with institutional data protection policies.

Instruments

- Standardized listening assessment: An institutionally validated test aligned with CEFR B1–B2, comprising multiple formats (gist/detail, inference under noise, accent-diverse segments). Parallel forms were equated via item response theory (IRT) during pilot.
- Standardized speaking assessment: A rater-moderated performance task battery (dialogic role-play, information-gap negotiation, and problem-solving discussion). Audio was rated holistically and analytically on CEFR-referenced scales for comprehensibility, prosody/fluency (speed, breakdown, repair), lexical–grammatical range/accuracy, and interactional competence (turn management, repair, responsiveness).
- Process-tracing analytics: Clickstream logs (caption toggling, density changes, pausing/replay timing), time-on-task, and interactive metrics (turn length, overlap, repair initiation types, response latency) extracted from the platforms.
- Observation protocol: A structured instrument capturing instructional fidelity (exposure to supports, timing of fades), task ecology (information gaps, negotiation opportunities), and learner behaviors (strategy prompts uptake), with low-inference indicators and field notes.
- Semi-structured interview guides: Separate protocols for learners and teachers exploring perceptions of task authenticity, regulation of supports, anxiety, perceived fairness of AI feedback, transfer across genres, and equity concerns (accent inclusivity).
- Questionnaires: (a) Metacognitive Awareness of Listening and Speaking Questionnaire (MALSQ), adapted and validated for dual-skill regulation with subscales for planning, monitoring, evaluation; (b) Foreign Language Classroom Anxiety Scale—modified oracy version (FLCAS-O); (c) Technology Acceptance and Trust in Feedback (TATF) scale tailored to AI-mediated oracy feedback. All instruments underwent cognitive interviewing

and psychometric evaluation (confirmatory factor analysis; reliability).

Listening assessments

Listening tests included four modules: (1) authentic news/audio essays (accent diversity; controlled lexical density), (2) conversational dialogues with overlapping speech and reduced forms, (3) lecture excerpts with visual reference cues, and (4) noisy environments (café hum, mild reverb). Items probed gist, local detail, referent tracking, discourse inference, and tolerance of ambiguity under time pressure. Audio parameters were standardized (RMS normalization; calibrated SNR for noise modules). Two equated forms (A/B) were assigned pre/post counterbalanced by section. IRT (2PL) and differential item functioning checks ensured comparability across conditions and demographics.

Speaking assessments

Speaking tasks mirrored classroom genres and elicited interactional moves: (a) role-play with information asymmetry (clarification, confirmation checks expected), (b) collaborative problem-solving requiring stance alignment and negotiation, and (c) short "explain-and-check" tasks to elicit comprehension checks and repair. Performances were recorded in small groups (triads) and in dyads for role-plays. Rating employed a four-dimension analytic rubric (comprehensibility; prosody/fluency; linguistic control; interactional competence). Two trained raters scored all samples; 20% were double-coded for interrater reliability (target ICC≥.80). Raters used a benchmarked video bank and underwent calibration with Many-Facet Rasch Measurement (MFRM) to adjust for severity and task effects.

Observation protocols

We used event-based and time-sampling observation. Event-based coding captured occurrences of strategy prompts, negotiation episodes, and scaffold fades; time sampling (5-minute intervals) noted task phase, tool use, and teacher mediation. Fidelity checklists documented adherence to condition-specific procedures (e.g., in C2, explicit plan–monitor–evaluate cycles; in C3/C4, adaptive parameter settings and feedback opportunities). Inter-observer agreement was calculated on 25% of sessions (Cohen's $\varkappa$ target ≥.75). Observers were blind to study hypotheses.

Semi-structured interviews

Learner interviews (45–60 minutes) occurred mid-intervention (Week 6) and post-intervention (Week 12) to capture evolving regulation and perceptions. Protocols probed: decision rules for caption density and pausing; dealing with accent/noise; experiences initiating/handling repair; reactions to AI feedback (usefulness, fairness, anxiety); and perceived transfer to non-classroom contexts. Teacher interviews focused on orchestration strategies, equity considerations, and assessment feasibility. Interviews were audio-recorded, transcribed verbatim, member-checked for factual accuracy, and anonymized.

Questionnaires

The MALSQ comprised 24 items (Likert 1–6) across planning (e.g., goal-setting, preview strategies), monitoring (attention shifts, comprehension checks), and evaluation (post-task analysis, transfer planning) for both listening and speaking. The FLCAS-O (20 items) targeted situational anxiety in dialogic contexts (e.g., fear of repair, live response pressure). The TATF (16 items) measured perceived legitimacy, interpretability, and trust in AI feedback. Instruments were administered pre (Week 1), mid (Week 6), and post (Week 12). CFA supported three-factor structures (CFI≥.95; RMSEA≤.06); omega reliabilities ranged .82–.90.

Data collection procedures

Week 0–1: Consent, orientation, baseline survey, and pretests (listening Form A/B counterbalanced; speaking task set 1). Rater calibration and platform onboarding included ethical use of analytics and privacy briefings.

Weeks 1–12: Instruction delivered per condition. Platform logs collected automatically; observation teams attended two sessions per section per week on a rotating basis (minimum 6 observations/section). Midpoint interviews and surveys occurred in Week 6; a low-stakes

formative assessment ensured continued engagement.

Week 12–13: Posttests (alternate listening form; speaking task set 2 with parallel difficulty); final surveys; exit interviews. Debriefing sessions clarified data use and provided aggregated feedback to classes.

Quantitative analysis

- Preliminary checks: Baseline equivalence across conditions (ANOVA/χ^2); missing data patterns (Little's MCAR test). Multiple imputation addressed sporadic missingness (<5%).
- Outcome modeling: Multilevel models (MLM) with learners nested in classes estimated condition effects on listening and speaking posttest scores, controlling for pretest scores, program, and instructor. Random intercepts for class and teacher accounted for clustering. Effect sizes were reported as adjusted Hedges g with 95% CIs.
- Process–outcome mediation: Multilevel structural equation modeling (MSEM) tested whether changes in process indicators (e.g., average caption density at second pass; strategic pause placement at clause boundaries; repair initiation frequency and timing; response latency) mediated condition effects on outcomes, controlling for anxiety and metacognitive gains.
- Speaking ratings calibration: MFRM adjusted for rater severity and task difficulty; resulting fair-average measures served as dependent variables in MLM.
- Questionnaire modeling: Latent change score models assessed growth in metacognitive awareness and anxiety. Cross-lagged paths explored directional relations between regulation, anxiety, and performance.
- Sensitivity analyses: Per-protocol versus intent-to-treat comparisons; robustness to ICC variation; subgroup analyses by baseline proficiency band (B1 vs. B2).

All analyses were conducted in R (lme4, lavaan) and Facets for MFRM. Assumptions (normality, homoscedasticity, linearity) were checked; robust SEs used where appropriate. Alpha was set at .05 with Benjamini–Hochberg correction for multiple comparisons in secondary outcomes.

Qualitative thematic analysis

We employed reflexive thematic analysis (Braun & Clarke, 2021) with abductive coding that combined sensitizing concepts from the theoretical framework (e.g., plan–monitor–evaluate; negotiation/repair; comprehensibility orientation) and inductive patterning from the data. The coding cycle:

1. Familiarization: Independent reading of transcripts and observation notes by three researchers.
2. Initial coding: Low-inference codes (e.g., "caption fade rationale," "repair avoidance," "feedback fairness concern").
3. Theme development: Aggregation into candidate themes (e.g., "strategic externalization," "affective buffering vs. dependence," "interactional timing as teachable target").
4. Review and refinement: Cross-case comparison by condition; negative case analysis to challenge early interpretations.
5. Integration: Joint displays aligned qualitative themes with process metrics (e.g., learners who articulated "timed pausing" also showed reduced subclausal pauses in logs).

Trustworthiness strategies included analyst triangulation, member checks on interpretive summaries, and an audit trail documenting coding decisions. NVivo supported data management.

Reliability and validity

- Instrument reliability: Internal consistency (ω) for questionnaires; test–retest stability for baseline measures (pilot). Listening test equating ensured parallel-form reliability ($r \approx .82$). Speaking ratings achieved interrater ICC≥.84 after MFRM calibration.
- Construct validity: CFA supported factor structures; convergent validity evidenced by correlations between metacognitive scales and

process indicators (e.g., planning scores with pre-task preview behaviors). Criterion validity for speaking included associations between interactional competence scores and independent listener comprehensibility judgments on a subset.

- Internal validity: Quasi-experimental risks mitigated by stratified random assignment, baseline covariates, fidelity monitoring, and MLM accounting for clustering. History and maturation threats were minimized by concurrent implementation across sections; testing effects reduced via parallel forms.

- External validity and ecological validity: Intact classes, authentic tasks, and routine institutional instruction bolster generalizability to similar tertiary EFL contexts. Transfer was probed through genre-diverse assessments.

- Reliability of observations and coding: Inter-observer agreement $\varkappa \geq .78$ on event codes; qualitative coding conducted by multiple analysts with periodic calibration meetings.

Ethical considerations

Institutional review board approval was obtained prior to data collection. Participants provided informed consent with options to opt out of specific data modalities (e.g., clickstream) without grade penalty. Data minimization was practiced: logs were pseudonymized; audio files were stored on encrypted servers; only de-identified aggregates were used for analysis. AI tools were configured to avoid storing raw speech beyond session-level analytics; no biometric or biometric-adjacent features were collected. Equity safeguards included: (a) accent-diverse input libraries; (b) feedback tuned to comprehensibility and interactional adequacy rather than nativeness; and (c) transparent communication about the probabilistic nature of automated feedback. Debriefings offered participants access to personal progress dashboards and guidance for continued learning. Teacher workload and wellbeing were considered in scheduling observations and providing planning templates to reduce orchestration burden.

Justification of methodological choices

- Quasi-experimental, cluster-based design reflects the realities of institutional scheduling and reduces contamination, while maintaining reasonable causal leverage with baseline controls and fidelity checks.

- Convergent mixed methods directly address the study's dual focus: estimating effects and explicating mechanisms. Process analytics operationalize theoretical constructs (e.g., metacognitive regulation; interactional moves), enabling mediation tests rarely feasible in classroom research.

- Speaking assessment emphasizes interactional competence and comprehensibility, aligning measurement with communicative goals and recent validity arguments in oracy assessment.

- Adaptive, genre-diverse listening tests with controlled parameters allow us to test the design principle of "calibrated difficulty with scaffold fading" and to examine transfer to noisy/novel conditions.

- Ethical orientation is integral, not adjunct: by configuring AI feedback to prioritize intelligibility and fairness and by curating accent diversity, we embody sociocultural commitments while leveraging technological affordances.

Limitations and mitigation

As a quasi-experiment in intact classes, residual selection effects cannot be fully excluded; however, baseline equivalence checks, MLM controls, and sensitivity analyses mitigate bias. Instructor effects remain possible; we addressed this via common training, shared materials, and random effects in modeling. Hawthorne effects may inflate engagement; process measures across weeks allow trend analysis beyond novelty spikes. Finally, platform dependence may limit replicability; we therefore document parameter settings and provide open task and rubric specifications to support adaptation.

Discussion

This study set out to examine whether coupling metacognitive scaffolding with AI-adaptive input and dialogic tasks would remediate persistent

bottlenecks in listening and speaking, and whether resulting gains would be traceable to observable process shifts (e.g., clause-aligned pausing, timely repair). The pattern of findings—largest improvements for the combined condition (C4), intermediate gains for metacognition-only (C2) and AI-only (C3), and smallest for the CLT baseline (C1)—supports an integrated account of oracy in which input processing, production, and regulation function as a coupled system. In what follows, we interpret the results through the proposed theoretical framework, situate them within recent international scholarship, and unpack implications for pedagogy, assessment, and technology design, with attention to sociocultural and contextual contingencies.

Interpreting outcomes through the integrated framework

The additive advantages observed in C4 align with the proposition that listening improvement requires not merely "more comprehensible input" but input tuned to cognitive constraints and embedded within tasks that demand response and enable repair (Field, 2021; Ellis, 2020). Krashen's emphasis on abundant, meaningful input explains little about when and how learners overcame specific decoding barriers such as reduced forms and overlapping talk. The process data—decreased subclausal pauses, increased clause-aligned pausing, and reduced response latency—are more parsimoniously explained by Cognitive Information Processing accounts of segmentation and working memory: supports that controlled speech rate and lexical density reduced extraneous load, while explicit prompting made boundary decisions an object of attention and practice. The observed fade in caption density in C3/C4 further suggests that scaffolds functioned as temporary external regulators whose withdrawal corresponded with internalized control, consistent with Metacognitive Awareness Theory (Goh & Vandergrift, 2022; Cross, 2023).

On the speaking side, the largest gains in interactional competence (timing, repair, responsiveness) in C4 indicate that output operated as hypothesis-testing in Swain's sense, but only when output was situated in negotiated, feedback-rich interaction (Swain's Output Hypothesis + Interaction Hypothesis). In the CLT baseline, exposure and practice led to modest improvements, yet avoidance of repair and delayed turn entry persisted—an instance of "speaking-as-display" without proceduralization of interactional moves. In C3/C4, conversational analytics and rater-aligned feedback redirected attention from mere accuracy to comprehensibility, prosody, and repair timing—features shown to drive listener judgments (Saito, 2021; Plonsky & Tavakoli, 2021). This reflects a design move grounded in our framework: alter the performance target to interactional adequacy and provide feedback at the granularity learners can act on within turns.

The mediating role of monitoring and process shifts speaks directly to the regulation layer: metacognitive gains (especially monitoring) partially accounted for achievement improvements, and clause-aligned pausing and repair initiation served as proximal conduits to comprehension and interactional competence. These patterns operationalize Vygotskian mediation: prompts, captions, and feedback acted as tools that organized attention and enabled performance in the Zone of Proximal Development; over time, such mediated routines became more autonomous, evidenced by declining scaffold reliance and more disciplined timing (Hall, Hellermann, & Pekarek Doehler, 2021). Constructivist commitments are visible in learners' reflective notes: decision rules and self-generated prompts indicate knowledge construction about when and how to deploy strategies.

Convergence and divergence with recent studies

Our findings agree with growing evidence that metacognitive instruction embedded in authentic tasks yields gains in listening and self-regulation beyond declarative strategy lessons (Goh & Vandergrift, 2022; Cross, 2023). The present study extends this line by quantifying enactment (e.g., placement of pauses) and connecting it to outcomes via mediation, addressing a common

critique that strategy research often infers enactment from self-report.

On multimodal support, the decline in caption density coupled with improved inference under noise aligns with research on graduated captioning and strategic fading (Montero Perez, 2022; Roquet et al., 2021). Where some studies report "caption dependence," particularly in lower-proficiency cohorts, our results suggest that dependence is not a necessary outcome: when caption use is explicitly planned, monitored, and faded—and when tasks require response rather than recall—learners can transition to less supported listening without loss of confidence.

Regarding speaking, we corroborate work showing that improvements in prosodic fluency and repair practices translate to higher comprehensibility more reliably than sheer speed or segmental correction (Saito, 2021; Plonsky & Tavakoli, 2021). Segmental accuracy gains were modest and non-differential across conditions, echoing studies in which segmental training outside interactional contexts yields limited communicative payoffs. By contrast, conversational feedback that targets phrasing and timing appears to facilitate perceptually salient improvements without overburdening learners' attentional resources.

Our AI-related results fit within a heterogeneous but promising literature. Studies have documented increased practice time and reduced apprehension with ASR and dialog systems, with variable effects on downstream communicative performance (Li & Hegelheimer, 2021; Park & Son, 2022). We add that AI's benefits are greatest when (a) difficulty is adaptively calibrated along speech rate and lexical density, (b) feedback prioritizes comprehensibility and interactional alignment, and (c) metacognitive scaffolding structures the use of these tools. This helps explain contradictions in prior work: systems optimizing for isolated accuracy metrics (e.g., word error rate) may yield shallow gains or even negative washback on interactional competence (Lai & Zheng, 2023). By reorienting feedback to interactionally relevant features and embedding it in dialogic tasks, we observed meaningful changes in turn timing and repair—effects that are less likely when AI is used as a decontextualized tutor.

Implications for listening pedagogy

Four implications emerge for the design of listening instruction. First, instruction should target segmentation as a teachable behavior under authentic temporal constraints. Our clause-aligned pausing results demonstrate that timing can be coached: learners can be trained to locate and exploit syntactic and prosodic boundaries to manage cognitive load, rather than pausing reactively within words or phrases. This suggests practical routines: brief pre-task drills that prime boundary cues, in-task prompts that cue pausing after conjunctions or intonation contours, and post-task reviews that visualize pause placement alongside transcripts. Such routines respect Cognitive Information Processing constraints and convert them into concrete pedagogic levers (Field, 2021).

Second, graduated captions should be treated as a strategic resource rather than an all-or-nothing accommodation. The decline in second-pass caption density, coupled with improved performance, supports protocols that (a) establish default caption-off first passes for gist, (b) permit low-density captions for detail checks, and (c) require reflective justification for any increase in support. Teachers can scaffold these decisions with classwide norms and individual plans, moving caption use from an undisciplined crutch to a deliberate tool (Montero Perez, 2022).

Third, authentic variability (noise, accents, overlapping talk) should be staged rather than suppressed. The largest item-level gains were in inference under noise. Rather than sanitizing input, teachers can manage variability through progression: begin with slightly slowed but authentic speech, add predictable noise layers, and gradually introduce accent diversity and overlap, all while sustaining communicative purpose. The goal is resilience, not perfection.

Fourth, listening tasks should culminate in response, not just recall. Tasks that require learners to act on what they heard—summarize

for a partner, ask a confirmation question, or problem-solve—reframe listening as preparation for speaking. This reframing grounds Swain's "push" to articulate and test understanding, making comprehension accountable to interactional success rather than item-level correctness.

Implications for speaking pedagogy

For speaking, the data counsel a shift in targets, tasks, and feedback. Target-wise, comprehensibility and interactional competence should anchor assessment and feedback. Our raters' comments and prosodic indices indicate that smoother phrasing and timely repair altered how speech was received even when segmental issues remained. Teachers can therefore prioritize phrasing at clause boundaries, intonation for new vs. given information, and explicit training in repair sequences (open-class repair initiators, partial repeats with question intonation), which our learners adopted more frequently in C3/C4.

Task-wise, monologic prompts should be reclustered into dialogic cycles that require listeners to shape speaker choices and speakers to manage co-construction. Information gaps, perspective shifts, and joint decision-making reliably elicited the repair and alignment moves that improved most.

Feedback-wise, conversation-analytic micro-feedback—short, specific comments tied to a turn ("good use of partial repeat before asking for detail," "pause before the reason clause to help your listener")—appears more actionable than broad accuracy notes. When augmented with visualizations (e.g., pause maps or pitch tracks) learners can see timing and phrasing as manipulable, not mysterious.

Implications for communicative competence and oracy

Our results substantiate a view of communicative competence that foregrounds co-construction: listeners anticipate and segment so that they can respond; speakers monitor for uptake cues and adjust turns; both participants negotiate alignment. Gains concentrated in interactional competence—the very dimension that traditional curricula and assessments often sideline. To the extent that oracy is adopted as a curricular priority, assessments must evolve. The Many-Facet Rasch calibration of our speaking ratings proved critical: without rater and task adjustments, interactions between feedback type and performance may be obscured. Program-level reforms might include integrating interactional measures (repair initiation, response latency to question type, backchannel appropriateness) into routine assessment, providing shared reference points that encourage classroom focus on dialogic skills (Harding & Brunfaut, 2020; Hall et al., 2021).

Digital learning environments: affordances and constraints

The digital environment in this study functioned as both a laboratory and a classroom. On the affordance side, it (a) delivered abundant, parameterized input; (b) captured process data granular enough to test theory; and (c) extended formative assessment bandwidth. Importantly, we observed no differential harm from routine technical glitches because instruction anticipated failure modes (offline backups, asynchronous alternatives). On the constraint side, orchestration load is real: teachers reported cognitive strain in timing scaffold fades and tracking multiple backchannels during live sessions. This implicates design for teacher mediation, not just learner practice. Dashboards that (a) surface only key indicators (e.g., average caption density trend; pause alignment outliers), (b) align terminology with rubrics, and (c) suggest next-step prompts can reduce load. Additionally, notifications must be managed: distraction was a recurrent complaint early in the term. Simple institutional norms (devices on do-not-disturb during listening cycles; negotiated pause windows) may be necessary complements to platform design (Hampel, 2021; Jin, 2022).

AI-assisted language education: promise, boundaries, and ethics

The AI layer delivered meaningful benefits when shaped by communicative goals and regulatory scaffolds. Three design features appear pivotal. First, adaptive difficulty calibrated along speech

rate and lexical density maintained challenge without overloading—an operationalization of i+1 that is measurable and adjustable. Second, feedback targets mattered: comprehensibility, phrasing, and repair alignment are closer to what interlocutors value than word-level accuracy. Third, transparency reduced anxiety and resistance: learners' trust improved when they saw how AI judgments mapped to human ratings and when feedback explanations referenced concrete turn features. These observations align with calls to reorient AI from grading to guidance and from correctness to communicative adequacy (Cukurova, Kent, & Luckin, 2022; Lai & Zheng, 2023; Huang & Chen, 2024).

At the same time, ethical concerns persist. Accent sensitivity remains an unsettled issue; even with configuration toward intelligibility, learners questioned fairness. Programmatic responses include (a) curating diverse input that normalizes multiple standards, (b) calibrating models on accent-diverse training data where possible, and (c) explicitly discussing with students how intelligibility and identity intersect. There is also a danger of metric myopia: when a dashboard quantifies what is easy to measure, teachers may inadvertently narrow instruction. Our mixed-methods approach—triangulating automated metrics with human ratings and qualitative accounts—offers a guardrail.

Sociocultural and contextual dimensions

Learning is situated. Instructors reported that norms around face and error display shaped willingness to initiate repair; students noted that "caption anxiety" spiked when fades were mandated without negotiation. The sociocultural lens suggests two adjustments. First, cultivate classroom norms that frame repair as competence, not failure. Modeling by teachers and explicit valuation of repair in rubrics counteracts face-threat perceptions (Kecskes, 2022). Second, co-design fading criteria: when learners participate in setting thresholds (e.g., "two consecutive 80% comprehension probes before caption reduction"), autonomy and buy-in increase, reflecting SCT's emphasis on mediated agency.

Context also includes institutional assessment and workload. Without alignment, classroom innovation can be nullified by summative washback. Where possible, departments should pilot oracy rubrics compatible with course tasks and provide credit for strategy evidence (e.g., reflection artifacts linked to process traces). On workload, orchestration demands professional learning: micro-credentialing in reading analytics, designing negotiation-rich tasks, and managing multimodal interaction would support scalability.

Explaining agreements and contradictions

Agreements with the literature—on metacognitive benefits, graduated captioning, and the primacy of prosodic/interactional features—stem from conceptual convergence: these lines of work share a view of oracy as real-time, resource-limited, and co-constructed. Apparent contradictions—e.g., reports of caption dependence or minimal AI effects—are reconcilable when design specifics are considered. Caption dependence emerges where fading is absent and tasks emphasize recall over response; AI null effects appear where systems optimize decontextualized accuracy or where feedback is opaque or normatively biased. The present results suggest that design decisions, not technology per se, determine outcomes.

Limitations shaping interpretation

While not the focus of a Discussion's implications, interpretive caution derives from design boundaries. As a quasi-experiment in intact classes, residual confounds cannot be fully eliminated despite baseline equivalence and multilevel controls. We focused on B1–B2 learners in a tertiary setting; generalization to beginners or advanced learners requires care. Additionally, segmental pronunciation did not change markedly; interventions emphasizing segmentals within the same integrated design could clarify how to support learners whose intelligibility is segmentally constrained. Finally, although anxiety decreased more in AI-enabled conditions, a subset of learners remained uneasy with recording or algorithmic judgment; multiple feedback modalities and opt-in privacy configurations are necessary to avoid exclusion.

Implications for research

Three research directions follow. First, mechanism testing: randomized experiments that manipulate single design levers (e.g., pause prompts vs. none; caption density schedules) within otherwise identical tasks would sharpen causal claims about specific supports. Second, transfer and durability: delayed posttests and out-of-class performance captures (e.g., workplace simulations) would assess whether gains persist and generalize. Third, equity audits: stratified analyses by L1 and accent, along with bias probes in AI feedback, are needed to move beyond ethical aspirations to empirical guarantees (Mercer, Warwick, & Ahmed, 2022; Kecskes, 2022).

Programmatic recommendations

- Curriculum: Embed listening-for-response and speaking-for-alignment cycles as weekly routines; sequence variability and scaffold fading across terms.
- Assessment: Adopt analytic rubrics that weight interactional competence and comprehensibility; incorporate process evidence (e.g., pause alignment metrics) into formative assessment.
- Teacher support: Provide planning templates for scaffold timing, exemplar negotiation sequences, and dashboard literacy; schedule collaborative moderation of speaking assessments to stabilize standards (Harding & Brunfaut, 2020).
- Technology procurement: Select tools that expose adjustable difficulty parameters, offer interpretable feedback aligned to interactional targets, and allow export of process traces for pedagogic reflection rather than surveillance.
- Learner development: Teach decision rules for supports; normalize and celebrate repair; cultivate reflective artifacts linking process to outcome (e.g., annotated transcripts with pause/repair marks).

Conclusion

This study examined whether an integrated pedagogy—combining adaptive, authentic input; dialogic, task-based output; and explicit metacognitive regulation—could address entrenched difficulties in L2 listening and speaking. Across standardized outcomes, process analytics, and learner–teacher accounts, the combined condition generated the largest gains in listening comprehension (especially inference under noise) and in speaking interactional competence (timely turn entry, repair, and prosodic fluency), with metacognition-only and AI-only variants yielding intermediate benefits. Anxiety declined most where adaptive tools and regulatory scaffolds were used together, and mediation analyses tied improvements to observable shifts in process (e.g., clause-aligned pausing, reduced response latency) and to growth in monitoring. These results support a coupled view of oracy: listening and speaking improve when input processing, production, and regulation are engineered to work in concert within communicative tasks.

Theoretical contributions

First, the findings refine how "comprehensible input" can be operationalized for listening: not as a static i+1, but as parametric tuning of speech rate, lexical density, and variability, with scaffolds that fade as control emerges (Field, 2021; Montero Perez, 2022). Second, they substantiate Swain's claim that output drives development by pushing hypothesis formation and control, but only when output is embedded in negotiation and supported by feedback aimed at comprehensibility, phrasing, and repair (Saito, 2021). Third, they instantiate Vygotskian mediation in digital contexts: captions, prompts, and AI feedback operated as tools in the Zone of Proximal Development, shifting from external regulation to internalized routines as learners adopted decision rules (Hall, Hellermann, & Pekarek Doehler, 2021). The framework thus integrates Input and Output Hypotheses, Interaction Hypothesis, cognitive load constraints, metacognitive regulation, CLT/TBLT, and constructivist task design into a single explanatory model of oracy as real-time, co-constructed action.

Methodological contributions

Methodologically, the study advances mixed-methods practice by:

- Translating theoretical constructs into measurable process indicators—pause alignment, caption-density trajectories, repair initiation timing—and modeling them as mediators of outcome change. This addresses a known weakness in strategy research, which often infers enactment from self-report (Goh & Vandergrift, 2022; Cross, 2023).
- Calibrating speaking ratings with Many-Facet Rasch Measurement to separate rater and task effects, aligning quantitative outcomes with interactional constructs emphasized in contemporary assessment validity arguments (Harding & Brunfaut, 2020).
- Implementing equated, genre-diverse listening forms with controlled variability (accent, noise, overlap), allowing inferences about resilience under authentic conditions rather than sanitized comprehension.

The convergence of standardized gains, process shifts, and qualitative accounts strengthens internal validity and provides a replicable template for classroom-embedded evaluation of complex interventions.

Pedagogical contributions

Practically, the study offers design-ready principles. For listening: teach segmentation as a skill; orchestrate graduated captions and strategic pausing; stage variability rather than suppress it; and make listening accountable to response. For speaking: target comprehensibility and interactional competence; design tasks that guarantee negotiation opportunities; and deliver micro-feedback on phrasing, timing, and repair. For regulation: embed plan–monitor–evaluate cycles tied to traceable behaviors, converting scaffolds into learner-authored decision rules. These principles align pedagogy with what interlocutors perceive and with how cognition operates under time pressure, moving instruction beyond "more practice" to "better-engineered practice."

Limitations

Four limitations shape inference. First, quasi-experimental assignment at the class level, though stratified and modeled with multilevel controls, cannot fully eliminate selection bias. Second, the context—B1–B2 university learners—limits generalizability; beginners and advanced learners may require different parameter ranges or supports. Third, segmental pronunciation gains were modest; learners whose intelligibility is segmentally constrained might benefit from targeted segmental modules integrated into the same interactional and metacognitive frame. Fourth, while anxiety decreased overall, a subset remained uneasy with recording and algorithmic judgment; offering multiple practice and feedback modalities is essential to avoid excluding risk-sensitive learners.

Future research

Two strands merit priority. Mechanism testing should use randomized component manipulations—e.g., presence/absence of pause prompts or specific caption schedules—within identical task ecologies to isolate causal levers and dose–response relations. Durability and transfer require delayed and out-of-class assessments, including workplace or community simulations, to evaluate whether gains in timing, repair, and resilient listening persist and travel across genres and modalities. Equity-focused studies should audit AI feedback for accent bias, stratify effects by L1 and identity variables, and test mitigations (e.g., intelligibility-oriented targets, accent-diverse calibration). Finally, teacher orchestration research is needed: design-based iterations clarifying which dashboard indicators are necessary and sufficient for day-to-day decision-making, and how professional learning can stabilize interactional assessment at scale.

Future developments in AI-supported language learning

The trajectory of AI in oracy education should move from correctness to communicative adequacy, from black-box scores to actionable explanations, and from static difficulty to pedagogically principled adaptivity (Cukurova, Kent, & Luckin, 2022; Lai & Zheng, 2023; Huang & Chen, 2024). Three priorities follow:

- Objective alignment: Train models on human judgments of comprehensibility and interactional quality, not only word error rates; expose feedback at the turn and phrase level, referencing timing and prominence patterns learners can change within a session (Saito, 2021).
- Pedagogic transparency: Provide rationales that map feedback to examples and to rubric language; allow teacher overrides and learner challenges, promoting trust and calibrating expectations.
- Ethical robustness: Curate accent-diverse training sets; conduct differential performance audits; and enable user control over target intelligibility profiles. Such moves align SCT's mediation ethic with inclusive communicative goals.

Originality and scholarly maturity

The study's originality lies in its concrete coupling of a theoretically plural model of oracy with a measurement strategy that makes regulation and interaction visible—and consequential—for learning. Rather than juxtaposing input, output, and interaction as competing claims, the design treats them as interdependent constraints and affordances that can be tuned in tandem, yielding measurable changes in the micro-timing of listening and speaking. By showing that clause-aligned pausing, repair initiation, and caption fading mediate achievement, the research moves beyond "what works" toward "how it works," offering a portable grammar of design for classrooms and tools alike. In a field often split between communicative ideals and techno-solutionism, this synthesis demonstrates that principled engineering of tasks, supports, and feedback can honor both cognitive realities and sociocultural ethics, delivering durable, transferable gains in oracy for multilingual participation.

References

Alexander, R. (2020). Oracy, education and social justice.

Boers, F. (2021). Evaluating multiword unit instruction in L2 speech.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis. Qualitative Research in Psychology, 18(3), 328–352.

Burston, J. (2020). Twenty years of MALL project implementation.

Carless, D. (2022). TBLT in East Asian school contexts revisited.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). Sage.

Cross, J. (2023). Metacognitive instruction for L2 listening: From awareness to enactment. Language Teaching Research, 27(6), 1243–1265.

Cukurova, M., Kent, C., & Luckin, R. (2022). Artificial intelligence in education: Promises and implications. British Journal of Educational Technology, 53(4), 1–18.

Cutler, A. (2021). Native listening: Language experience and speech perception.

Ellis, R. (2020). Task-based language teaching: Sorting out the misunderstandings. International Journal of Applied Linguistics.

Fang, F. (2022). Global Englishes and learner identities.

Field, J. (2021). Revisiting listening: Pedagogic priorities and processing constraints. Language Teaching, 54(4), 461–480.

Gao, C., Wang, Y., & Jiao, L. (2023). Adaptive feedback in AI-supported L2 learning.

Goh, C., & Vandergrift, L. (2022). Teaching and learning second language listening: Metacognition in practice.

Hall, J. K., Hellermann, J., & Pekarek Doehler, S. (2021). Interactional competence in practice. Annual Review of Applied Linguistics, 41, 1–26.

Hall, J. K., Hellermann, J., & Pekarek Doehler, S. (2021). The Routledge handbook of interactional competence.

Hampel, R. (2021). Disruptive technologies and online language learning.

Harding, L., & Brunfaut, T. (2020). Testing speaking and listening: Validity and interactional perspectives. Language Testing, 37(4), 555–575.

Huang, Y., & Chen, C. (2024). Large language models for L2 speaking: Feedback, transparency, and ethics. System, 120, 102993.

Jiang, Y., & Dewaele, J.-M. (2020). Anxiety, enjoyment, and L2 performance.

Jin, L. (2022). Online interaction in L2 learning communities.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. Educational Researcher, 33(7), 14–26.

Kang, O., & Johnson, D. (2021). L2 speech comprehensibility and listener factors.

Kecskes, I. (2022). Intercultural pragmatics and English as a lingua franca.

Lai, C., & Zheng, D. (2023). Rethinking AI for language learning: From accuracy to communicative adequacy. Computer Assisted Language Learning, 36(7), 1471–1493.

Li, Z., & Hegelheimer, V. (2021). ASR-mediated feedback for L2 speaking.

Montero Perez, M. (2022). Captions, cognitive load, and L2 listening: When and how to fade support. Language Learning & Technology, 26(3), 1–26.

Morgan, D. L. (2014). Pragmatism as a paradigm for mixed methods research. International Journal of Multiple Research Approaches.

Palinkas, L. A., et al. (2015). Purposeful sampling for qualitative data collection and analysis. Administration and Policy in Mental Health.

Park, M., & Son, J.-B. (2022). Chatbots in language education: A critical review.

Plonsky, L., & Tavakoli, P. (2021). Fluency in L2 research and pedagogy.

Roquet, H., et al. (2021). Caption use and learning outcomes in video-based L2 input.

Saito, K. (2021). What makes L2 speech comprehensible? Studies in Second Language Acquisition, 43(5), 1205–1229.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs. Houghton Mifflin.

Tavakoli, P., & Wright, C. (2020). Task design and speech fluency.

Viberg, O., & Grönlund, Å. (2021). Mobile learning in higher education.