# Theories and Mechanisms for AI-Powered ESL Speaking System Design

---

**Student:** Joe (Beiqiao) Hu
**Course:** TIMG 5104 Directed Studies
**Date:** December 1, 2025

---

## Abstract

This scoping review systematically maps the theories and mechanisms that support the design of AI-powered speaking practice systems for adult English as a Second Language (ESL) learners. Following PRISMA-ScR guidelines, the review synthesized 17 empirical studies published between 2015 and 2025, selected through a mechanism-focused purposive sampling strategy employing an AI-human hybrid triangulation protocol. The synthesis identifies six foundational learning theories—including Skill Acquisition Theory, the Noticing Hypothesis, and Transfer-Appropriate Processing—that justify specific design choices. Ten instructional mechanisms were isolated, with evidence supporting a "phased practice architecture" that transitions from blocked to interleaved scheduling to optimize both fluency and transfer. Key findings indicate that explicit, multi-modal feedback (ASR for pronunciation, pending prompts for grammar, LLM dialogue for discourse) significantly outperforms single-mode correction. Furthermore, embedding AI tools within structured pedagogical frameworks (e.g., BOPPPS) amplifies their effectiveness by fostering metacognition. The report proposes four evidence-based design applications, including a "Theory-Mechanism-Design" (TMD) logic for feature validation, to guide the development of business-ready AI-ESL solutions.

# Executive Summary

## Purpose and Scope

This scoping review synthesizes 17 empirical studies (2015–2025) to map the theories and instructional mechanisms that should drive the design of AI-powered speaking practice systems for adult English as a Second Language (ESL) learners. Rather than cataloguing products, the review focuses on *why* specific interventions work and how to translate them into product-ready features. The primary design target is an AI-enabled speaking coach for real-world, impromptu fluency, where learners rehearse spontaneous speech, receive multi-layered feedback, and build long-term autonomy.

**Key outputs:** The review identifies **6 foundational learning theories**, isolates **10 instructional mechanisms**, and proposes **4 evidence-based design applications** with concrete KPIs, together forming a theory-driven blueprint for business-ready AI-ESL speaking systems.

## Methods

Following PRISMA-ScR guidance, the review used purposive, mechanism-focused sampling rather than exhaustive coverage. Across six databases (Scopus, Web of Science, EBSCO, JSTOR, ERIC, PsycInfo), an AI–human hybrid screening protocol processed 2,877 records and retained 17 mechanism-rich empirical studies that reported speaking outcomes for adult learners. The Three-Layer AI-Human Triangulation protocol replaced a second human screener with multi-model LLM consensus plus targeted human adjudication, allowing broader search coverage while maintaining transparency and auditability. Full search strings, platform counts, and AI prompts are documented in the methodological appendix of the main report.

## Key Theoretical Foundations

**Skill Acquisition Theory**: Speaking improves when learners move from declarative to procedural control through heavy, structured practice. Design implication: schedule high-volume, scaffolded practice with clear progression from accuracy to automaticity.

**Noticing Hypothesis**: Learners must detect gaps between their output and targets. Design implication: surface precise pronunciation, grammar, and discourse gaps with prompts that elicit self-repair before showing models.

**Transfer-Appropriate Processing**: Practice should mirror real-world use (timed, impromptu, varied topics) to generalize. Design implication: simulate performance contexts and mix task types once stability is reached.

**Usage-Based / Constructionist Learning**: Repeated use of high-value constructions drives fluency and complexity. Design implication: track and recycle signature phrases and sentence frames with user-facing dashboards.

**Sociocultural / Affective Theories**: Safe, scaffolded interaction lowers anxiety and raises willingness to communicate. Design implication: non-judgmental AI partners, clear goals, and supportive nudges before raising difficulty.

## Top 5 Instructional Mechanisms

1) **Phased practice:** start blocked, then interleave tasks as accuracy stabilizes to balance fluency gains and transfer.

2) **Adaptive spacing**: tune inter-session intervals based on performance to sustain retrieval and long-term retention.

3) **Explicit ASR pronunciation feedback**: immediate, segmental feedback with replay/compare to drive self-correction.

4) **Pending/elicited feedback**: delay full corrections to prompt self-repair, deepening processing of form.

5) **Structured session orchestration (e.g., BOPPPS):** each session has clear objectives, guided practice, feedback, reflection.

## Top 5 Design Principles for AI-ESL Systems

1) **Architect the practice curve**: progression engine governs blocked->interleaved and controlled->spontaneous tasks.

2) **Prioritize metacognition and self-repair**: prompts, hints, and reflection steps make learners active problem-solvers.

3) **Visualize invisible gains**: dashboards for construction growth, stability, and signature phrases.

4) **Lower anxiety before load**: safe practice space, confidence nudges, and optional warm-ups before timed runs.

5) **Hybridize feedback**: combine ASR (micro), pending grammar prompts (meso), and LLM discourse feedback (macro).

## Three Biggest Limitations of the Evidence Base

1. **Demographic narrowness:** ~2/3 of studies are East Asian university learners; generalizability remains limited.
2. **Short durations:** most interventions last 2-8 weeks with little delayed follow-up; durability of gains is inferred.
3. **Emerging mechanisms**: construction tracking and pending feedback rely on few studies; treat as high-potential hypotheses.

## Three Recommended Next Steps

1. Pilot the phased practice + tri-modal feedback stack for 4-6 weeks (N=15-20); capture pre/post pronunciation, fluency, discourse plus in-app logs.
2. Calibrate scheduling/feedback thresholds with early user data (A/B or bandits) to tune blocked->interleaved shifts, spacing, and feedback intensity.
3. Broaden recruitment beyond East Asia to test cultural moderators; report outcomes and anxiety patterns by cohort.

# 1. Introduction

## 1.1 Background and Scope

This review synthesizes theories and mechanisms to guide the design of AI-powered systems that support adult English as a Second Language (ESL) learners in developing speaking proficiency. The design targets several well-documented challenges for ESL learners:

1. **Lack of Practice Opportunities:** Insufficient time for meaningful, interactive speaking practice in typical classroom settings. Constraints such as limited in-class time and large class sizes make it impossible to provide sufficient individualized practice for each student (Mingyan et al., 2025).

2. **Speaking Anxiety:** Fear of negative evaluation and linguistic insecurity, which leads to reluctance to speak and hinders skill development. Foreign language anxiety significantly impacts learners' willingness to communicate and their oral proficiency development (Zheng et al., 2025).

3. **Inadequate Feedback:** The difficulty for human teachers to provide consistent, individualized, and immediate feedback required for effective learning. Teachers are unlikely to provide individual feedback to each student due to time constraints (Ngo et al., 2024; Sun, 2023).

## 1.2 Review Objective

This literature review synthesizes peer-reviewed empirical evidence to identify: (1) learning theories that support AI-powered speaking practice system design, (2) instructional mechanisms that effectively improve speaking outcomes in AI-powered environments, and (3) the differential impacts of AI technologies on pronunciation, fluency, and discourse competence. The findings will provide evidence-based foundations for designing an AI-powered speaking practice system for adult ESL learners.

## 1.3 Research Questions

The review addresses two primary research questions, structured using the Population-Concept-Context (PCC) framework:

- **Population:** Adult ESL learners in university or higher education contexts
- **Concept:** Learning theories, theoretical frameworks, cognitive and sociocultural models supporting oral skill acquisition
- **Context:** AI-powered speaking practice, digital language learning environments, oral proficiency development

**RQ1:** What learning theories (derived from empirical studies) support the design of AI-powered speaking practice systems for adult ESL learners?

**RQ2:** What instructional mechanisms (evidenced in empirical research) effectively improve speaking outcomes when implemented in AI-powered environments?

---

# 2. Methods

## 2.1 Literature Search Strategy

Scoping review with purposive, mechanism-rich sampling guided by PRISMA-ScR logic. Six databases (Scopus, Web of Science, EBSCO Education Source, JSTOR, ERIC, PsycInfo) were searched for 2015-2025 adult ESL/EFL speaking studies involving AI support or practice-scheduling mechanisms.

A multi-LLM + human triangulation workflow screened 2,877 records to 1,426 uniques, 128 full-texts, and 17 retained studies with speaking outcomes. Details of search strings, platform hit counts, prompts, and adjudication logs are now summarized in Appendix B (Detailed Search Protocol).

## 2.2 Inclusion and Exclusion Criteria

Inclusion: adult/higher-ed ESL/EFL learners; AI-supported or mechanism-focused speaking interventions; reported speaking outcomes (pronunciation, fluency, discourse, accuracy, affect); empirical designs; English; 2015-2025.

Exclusion: K-12 only; non-AI or unrelated interventions; no speaking outcomes; reviews/commentaries; non-English.

## 2.3 Study Selection Process

The study selection followed PRISMA with an AI–human hybrid screen. From 2,877 initial records (57 searches), 1,426 unique records remained after de-duplication. Title/abstract screening kept 128; full-text review retained 17 mechanism-rich studies on adult ESL speaking outcomes. AI model votes (multi-LLM) were compared and then human-adjudicated at each decision point. Detailed search strings, hit counts by platform, screening prompts, and adjudication logs are documented in Appendix B (Detailed Search Protocol and AI–Human Triangulation).

PRISMA flow diagram appears in Appendix A.1. Quality appraisal rubric resides in Appendix C.

---

# 3. Evidence Summary

## 3.1 Raw Evidence Highlights

**Table 1:** Summary of Primary Evidence from Included Studies

| Study | Design & Learners | Key Intervention | Primary Outcomes (Raw Evidence) |
|---|---|---|---|
| Li & DeKeyser (2019) | 80 adult L1-English learners of L2 Mandarin; 4 ISI × RI groups | Daily vs. weekly spacing for tonal production | Short ISI preserved procedural pronunciation; longer ISI aided declarative recall; RI drove forgetting |
| Suzuki (2017) | 60 learners in miniature language; 4 sessions | 3.3-day vs. 7-day spacing | Shorter ISI delivered higher post-test accuracy at 7 & 28 days |
| Suzuki (2021) | 68 Japanese undergrads; blocked vs. interleaved vs. control | 3-day narrative repetition | Blocked boosts articulation rate and mid-clause pause control; blocked advantages during training; transfer to novel prompts was limited |
| Suzuki & Hanzawa (2022) | 79 learners; massed, short-spaced, long-spaced, control | Same story repeated under different schedules | Massed sharply reduced pauses but slowed speech and spiked verbatim repetition |
| Suzuki, Eguchi, & de Jong (2022) | 50 learners; blocked vs. interleaved | Construction reuse analysis | Blocked practice induced higher lexical/POS reuse; reuse correlated with fluency gains in interleaved condition |
| Zhang, Yi, & Zhou (2023) | 90 freshmen; problem-solving tasks | Blocked vs. interleaved vs. control | Both experimental groups outperformed control; interleaved group was generally more advantageous than blocked, except for silent pause numbers |
| Tejedor-García et al. (2020) | 18 adult Spanish learners | CAPT tool vs. classroom minimal-pair training | CAPT users showed significantly larger pronunciation gains; CAPT intensity and automatic–human scoring alignment reported |
| Ngo, Chen, & Lai (2024) | Meta-analysis; 15 studies | ASR feedback typology | Overall effect g=0.69; explicit corrective feedback large (g=0.86); segmental effects large, suprasegmental small; medium–long duration > short; with peers > alone |
| Sun (2023) | 61 Chinese learners; 14-week course | ASR with peer correction | ANCOVA favored ASR+peer for accentedness, comprehensibility, |

| Study | Design & Learners | Key Intervention | Primary Outcomes (Raw Evidence) |
|---|---|---|---|
| | | vs. teacher feedback | spontaneous speech, and global speaking |
| Evers & Chen (2021) | 92 adult ESL learners; 12-week course | ASR + peer vs. ASR only vs. teacher | ASR+peer produced largest gains; learning style moderated effects |
| Mingyan et al. (2025) | 63 undergrads; 10-week after-class tasks | Liulishuo AI app + WeChat vs. WeChat | Experimental group outperformed on overall speaking, pronunciation, fluency ($p < 0.01$) |
| Farooqi (2025) | 44 IELTS candidates; 3 weeks | SmallTalk2Me AI feedback | IELTS speaking bands +12.12%; WPM +11.18%; learners reported lower anxiety |
| Abdelhalim & Alsehibany (2025) | 71 Saudi undergrads | ChatGPT + SpeechAce in class | Segmental accuracy improved; limited suprasegmental gains; motivation increased for Current L2 Self and Intended Learning Effort; Ideal L2 Self unchanged |
| Lai (2025) | 89 Taiwanese undergrads; 18 weeks | ChatGPT integrated into BOPPPS framework | ANCOVA showed significant gains across all speaking sub-skills ($\eta^2$ up to .105) |
| Yang et al. (2025) | 52 Chinese undergrads; 9 weeks | AI-supported interleaved training | EG exceeded CG on multiple rubric dimensions of impromptu speaking performance ($p<0.001$); engagement predicted speaking, with technology acceptance as a mediator |
| Zargaran (2025) | 60 IELTS learners; 30-hour course | Pending feedback vs. immediate correction | Grammatical Range & Accuracy improved; significant positive correlation between pending-feedback frequency and GR/A gains |
| Zheng et al. (2025) | 83 freshmen; 4-week trial | GPT-4 collaborative dialogue vs. control | LLM group improved oral proficiency, WTC, self-efficacy, and reduced anxiety |

## 3.2 Cross-Study Insights

- **Explicit, scaffolded feedback delivers larger gains:** Across ASR and pending-feedback studies, explicit cues or elicited self-repairs outperformed generic transcripts or immediate supply of answers (Ngo et al., 2024; Sun, 2023; Zargaran, 2025).

- **Timing matters:** Dense scheduling accelerates procedural speech; excessive massing harms speed, whereas interleaving or moderate spacing supports transfer (Li & DeKeyser, 2019; Suzuki, 2021; Zhang et al., 2023). Note: some metrics (e.g., silent pause counts) may not follow the same pattern (Zhang et al., 2023).

- **Social mediation amplifies AI benefits:** Peer collaboration or LLM partners added motivational and noticing advantages beyond solo AI loops (Evers & Chen, 2021; Zheng et al., 2025; Abdelhalim & Alsehibany, 2025).

- **Structured orchestration raises discourse quality:** Lesson frameworks (BOPPPS, interleaved routines) improved discourse management, confidence, and adherence (Lai, 2025; Yang et al., 2025).

## 3.3 Demographic Characteristics

Understanding the demographic composition of the included studies is essential for evaluating the generalizability of the identified mechanisms and assessing potential boundary conditions. The participant populations across the 17 studies exhibit notable patterns in terms of geographic origin, linguistic background, and educational context, which have important implications for the applicability of these findings to diverse ESL learning populations.

**Table 2:** Demographic Characteristics of Included Studies

| Study | N | Learner Background | Region |
|---|---|---|---|
| Li & DeKeyser (2019) | 80 | English L1 learners of Mandarin, adults | USA |
| Suzuki (2017) | 60 | Adult learners (miniature language study) | Japan |
| Suzuki (2021) | 68 | Japanese university students | Japan |
| Suzuki & Hanzawa (2022) | 79 | Japanese university students | Japan |
| Suzuki, Eguchi, & de Jong (2022) | 50 | Japanese university students | Japan |
| Zhang, Yi, & Zhou (2023) | 90 | Chinese freshmen | China |
| Tejedor-García et al. (2020) | 18 | Spanish adult learners | Spain |

| Study | N | Learner Background | Region |
|---|---|---|---|
| Ngo, Chen, & Lai (2024) | 15 studies | Meta-analysis (mixed populations) | Mixed |
| Sun (2023) | 61 | Chinese university students | China |
| Evers & Chen (2021) | 92 | Adult ESL learners (mixed L1) | USA |
| Mingyan et al. (2025) | 63 | Chinese undergraduates | China |
| Farooqi (2025) | 44 | Saudi IELTS candidates | Saudi Arabia |
| Abdelhalim & Alsehibany (2025) | 71 | Saudi undergraduates | Saudi Arabia |
| Lai (2025) | 89 | Taiwanese university students (Chinese L1) | Taiwan |
| Yang et al. (2025) | 52 | Chinese undergraduates | China |
| Zargaran (2025) | 60 | IELTS learners | Iran/Middle East |
| Zheng et al. (2025) | 83 | Chinese freshmen | China |

---

**Geographic Distribution:** - **East Asia:** 10 studies (China: 6, Japan: 4, Taiwan: 1) - **Middle East:** 3 studies (Saudi Arabia: 2, Iran: 1) - **North America:** 2 studies - **Europe:** 1 study (Spain) - **Meta-analysis:** 1 study (mixed populations)

The demographic analysis reveals substantial geographic concentration: nine of the sixteen primary studies (56%) focused on East Asian university learners, with six studies conducted in China, three in Japan, and one in Taiwan. Including studies conducted in East Asian contexts regardless of participant specification, ten studies (63%) were located in the region. This concentration limits the cross-cultural generalizability of the identified mechanisms, as all East Asian studies involved learners from Chinese or Japanese linguistic backgrounds. The predominance of university-level contexts (fifteen of seventeen studies) further suggests that findings may require adaptation for community-based, workplace, or self-directed learning environments. Future research should prioritize validation studies with learners from underrepresented regions, including sub-Saharan Africa, Latin America, South Asia, and Eastern Europe, to establish boundary conditions and identify necessary adaptations for diverse learner populations.

---

## 4. Theoretical Foundations

"Theory" refers to a principled explanation of how knowledge is acquired, retained, and transferred in second-language speaking. The six categories below emerge directly from the 17-study evidence base and anchor later mechanism choices.

### 4.1 Skill Acquisition Theory

Skill Acquisition Theory posits that learning a new skill, such as speaking a foreign language, progresses through three stages: declarative (understanding the rules), procedural (practicing the skill), and automatic (performing the skill without conscious thought). The transition between these stages is facilitated by practice and feedback. The reviewed studies by Li & DeKeyser (2019) and Suzuki (2017) demonstrate that the type and schedule of practice significantly impact the proceduralization and retention of language skills.

A system designed on this principle would be effective because it can provide the extensive, repetitive practice needed to drive skills from the declarative to the automatic stage—a volume of practice impractical in traditional classrooms.

**Evidence Anchors:** Li & DeKeyser, 2019; Suzuki, 2017; Suzuki, 2021; Suzuki et al., 2022; Suzuki & Hanzawa, 2022

### 4.2 Retrieval Practice and Desirable Difficulties

Effortful, spaced retrieval of information from memory is a powerful method for stabilizing that memory. An optimal Inter-Session Interval (ISI) depends on the Retention Interval (RI) and task complexity, creating a "desirable difficulty" that enhances long-term learning.

The proposed system's design is justified because it can manipulate practice schedules (spacing, interleaving) to create these desirable difficulties, optimizing retention.

**Evidence Anchors:** Li & DeKeyser, 2019; Suzuki, 2017; Suzuki & Hanzawa, 2022

### 4.3 The Noticing Hypothesis

According to the Noticing Hypothesis, learners must consciously notice a linguistic feature in the input for it to become part of their language acquisition process (Schmidt, 1990). Feedback is crucial for drawing learners' attention to the gap between their current proficiency and the target language. The research by Zargaran (2025) on "pending feedback" and the meta-analysis by Ngo et al. (2024) on explicit corrective feedback confirm that feedback mechanisms that enhance noticing are highly effective.

The proposed system would be effective because its feedback mechanisms can be engineered to make these gaps salient, forcing the learner to notice errors and initiating the process of correction and learning.

**Evidence Anchors:** Tejedor-García et al., 2020; Ngo et al., 2024; Zargaran, 2025

### 4.4 Transfer-Appropriate Processing (TAP)

This principle states that practice is most effective when it closely resembles the cognitive processes required by the target task. To get better at impromptu speaking, one must practice in conditions that demand similar flexibility and adaptability.

The proposed system would be effective because it can use mechanisms like interleaved practice to simulate the cognitive demands of real-world conversation, improving the transfer of skills to new and unseen prompts.

**Evidence Anchors:** Suzuki, 2021; Zhang et al., 2023; Yang et al., 2025

### 4.5 Usage-Based and Constructionist Learning

These theories propose that language is acquired by entrenching “constructions” (form-meaning pairings) through repeated use. Frequency of use strengthens memory traces and speeds up retrieval. The study by Suzuki et al. (2022) on construction reuse demonstrates how task repetition schedules can be designed to promote the entrenchment and proceduralization of grammatical patterns.

The proposed system’s design is justified because its core mechanisms, like task repetition, would systematically increase the frequency of use for specific linguistic constructions, leading to faster, more fluent speech.

**Evidence Anchors:** Suzuki, Eguchi & de Jong, 2022

### 4.6 Sociocultural and Affective-Motivational Theories

Sociocultural Theory emphasizes the importance of social interaction in learning (Vygotsky, 1978). Learning occurs within the Zone of Proximal Development (ZPD), where a learner, with the help of a “More Knowledgeable Other” (MKO), can achieve a higher level of performance than they would alone. In the context of an AI-ESL system, both AI tutors and peers can function as MKOs. Studies by Zheng et al. (2025) and Lai (2025) show that AI-powered dialogue partners can effectively scaffold learning and reduce anxiety.

The proposed system would be effective because it can act as a patient, non-threatening MKO, creating a low-anxiety practice environment that boosts confidence and engagement.

**Evidence Anchors:** Zheng et al., 2025; Lai, 2025; Sun, 2023; Evers & Chen, 2021

### 4.7 Priority Ranking of Theoretical Lenses

While this review identifies six complementary theoretical perspectives, not all exert equal influence on the design recommendations. Based on the density and strength of the empirical evidence mapped in this review, the following priority ranking is proposed:

**High-priority theories** include **Skill Acquisition Theory**, the **Noticing Hypothesis**, and **Transfer-Appropriate Processing (TAP)**. These frameworks are directly instantiated in multiple high-quality studies on task repetition, distribution of practice, and feedback, and they provide the primary justification for the phased practice architecture and tri-modal feedback loop at the core of the proposed system.

**Medium-priority theories** include **Usage-Based / Constructionist Learning** and **Sociocultural / Affective-Motivational Theories**. Usage-based perspectives are strongly supported by one mechanism-focused construction reuse study and conceptually align with the construction tracking dashboard, while sociocultural and affective perspectives are repeatedly invoked to explain reductions in anxiety and increases in willingness to communicate when AI partners or structured frameworks are used.

**Emerging theoretical refinements** center on **retrieval practice optimization for speech**, which integrates spacing, interleaving, and effortful retrieval into a unified perspective on how practice conditions should be tuned. Although the underlying cognitive science is well established, direct applications to AI-mediated speaking interventions are still limited to a small set of studies. For this reason, retrieval-based scheduling is treated as a promising design direction that should be empirically calibrated in future pilots.

---

## 5. Instructional Mechanisms

Analysis of the 17 studies identified 10 mechanisms across 3 categories. These mechanisms represent the HOW—the specific instructional strategies and technologies that operationalize the theories described in Section 4.

### 5.1 Practice and Scheduling Mechanisms

#### *5.1.1 Adaptive Spacing Between Sessions*

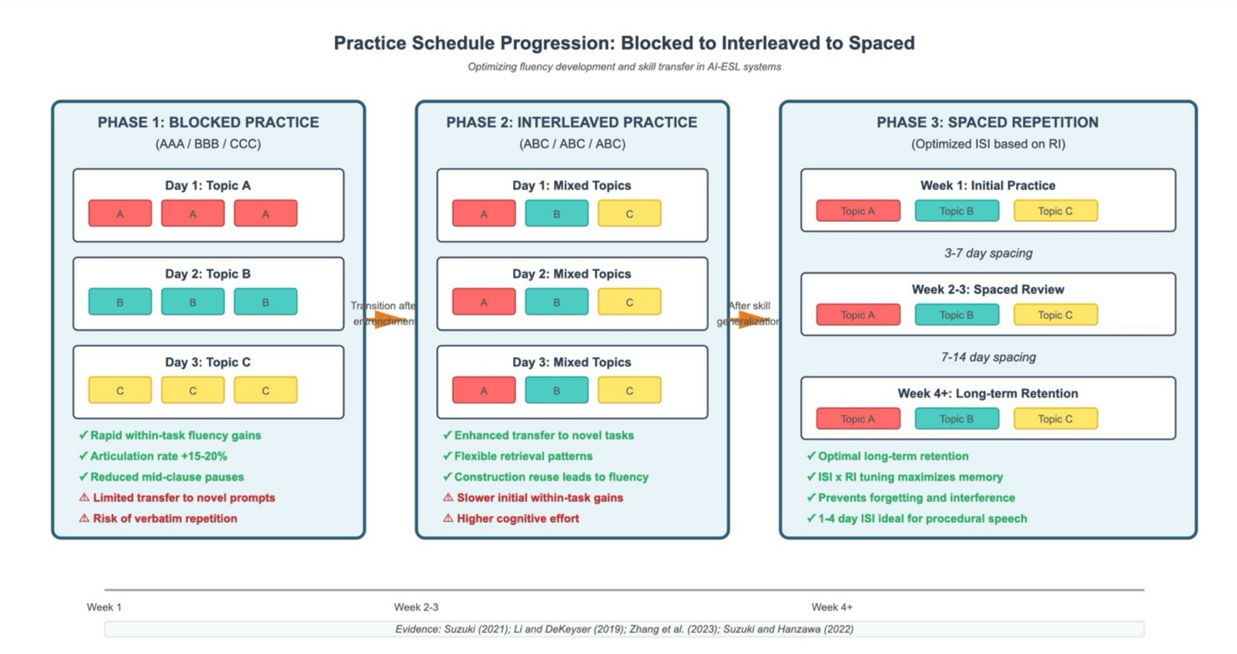
**Mechanism Description:** Distribute practice over time with inter-session intervals (ISI) tuned to retention goals and knowledge type—shorter ISI often favors proceduralized oral performance, while longer ISI can support declarative retention—rather than massing practice in a single session.

**Evidence:** Li & DeKeyser, 2019; Suzuki, 2017

#### *5.1.2 Task Repetition Schedules*

**Mechanism Description:** Use blocked cycles (AAA) for rapid within-task fluency. Use interleaved cycles (ABC) to promote transfer to new tasks. Avoid excessive massed practice, which can be a "double-edged sword," hurting speed and encouraging verbatim repetition even as it reduces pauses.

**Figure 1**: Practice Schedule Progression




*This figure illustrates the recommended progression from blocked practice (focusing on a single task type to build foundational fluency) through interleaved practice (mixing multiple task types to develop flexible retrieval and transfer) to spaced review (maintaining long-term retention through distributed rehearsal). The visual representation demonstrates how learners advance through increasingly complex scheduling patterns as their proficiency develops.*

**Evidence:** Suzuki, 2021; Zhang et al., 2023; Yang et al., 2025; Suzuki & Hanzawa, 2022

### *5.1.3 Construction Reuse with Variation*

**Mechanism Description:** Deliberately recycle target grammatical and lexical constructions across different prompts, while gradually adding variation and fading scaffolds.

**Evidence:** Suzuki, Eguchi & de Jong, 2022

## 5.2 Feedback Mechanisms

### *5.2.1 Explicit Corrective Feedback via ASR*

**Mechanism Description:** Provide immediate, explicit highlights of phoneme, stress, or timing errors with clear exemplars. A meta-analysis shows explicit feedback is more effective than indirect feedback (e.g., simple transcription).

**Evidence:** Ngo et al., 2024; Tejedor-García et al., 2020; Sun, 2023

### *5.2.2 Pending/Elicited Feedback*

**Mechanism Description:** For certain error types (especially grammatical), withhold the correct answer initially and instead provide prompts that cue the learner to self-repair, encouraging deeper cognitive processing and metacognitive awareness.

**Evidence:** Zargaran, 2025

## 5.3 Scaffolding and Support Mechanisms

### *5.3.1 Partner Scaffolding (ASR+Peer or LLM Partner)*

**Mechanism Description:** Augment solo ASR practice with either peer review or a collaborative dialogue with an LLM partner. This enhances noticing and lowers anxiety.

**Evidence:** Evers & Chen, 2021 (ASR+peer was most effective); Zheng et al., 2025 (LLM partner reduced anxiety)

### *5.3.2 Structured Lesson Orchestration (e.g., BOPPPS)*

**Mechanism Description:** Frame AI-supported tasks within a proven pedagogical model (Bridge-in, Objective, Pre-assessment, Participatory learning, Post-assessment, Summary).

**Evidence:** Lai, 2025

### *5.3.3 Mobile Micro-Practice with Analytics*

**Mechanism Description:** Assign short, frequent mobile practice sessions that provide automatic feedback, extending learning beyond the classroom.

**Evidence:** Mingyan et al., 2025; Farooqi, 2025

### *5.3.4 Graduated Difficulty and Mastery Gates*

**Mechanism Description:** Increase task difficulty (e.g., longer speaking windows, fewer prompts) once mastery thresholds are met; provide fallbacks if learners struggle.

**Evidence:** Suzuki, 2021; Suzuki & Hanzawa, 2022

### *5.3.5 Confidence Nudges and Reflection*

**Mechanism Description:** Incorporate brief pre/post self-assessments, success framing, progress visualizations, and reflective prompts to maintain motivation, build self-efficacy, and reduce anxiety.

**Evidence:** Zheng et al., 2025; Lai, 2025

---

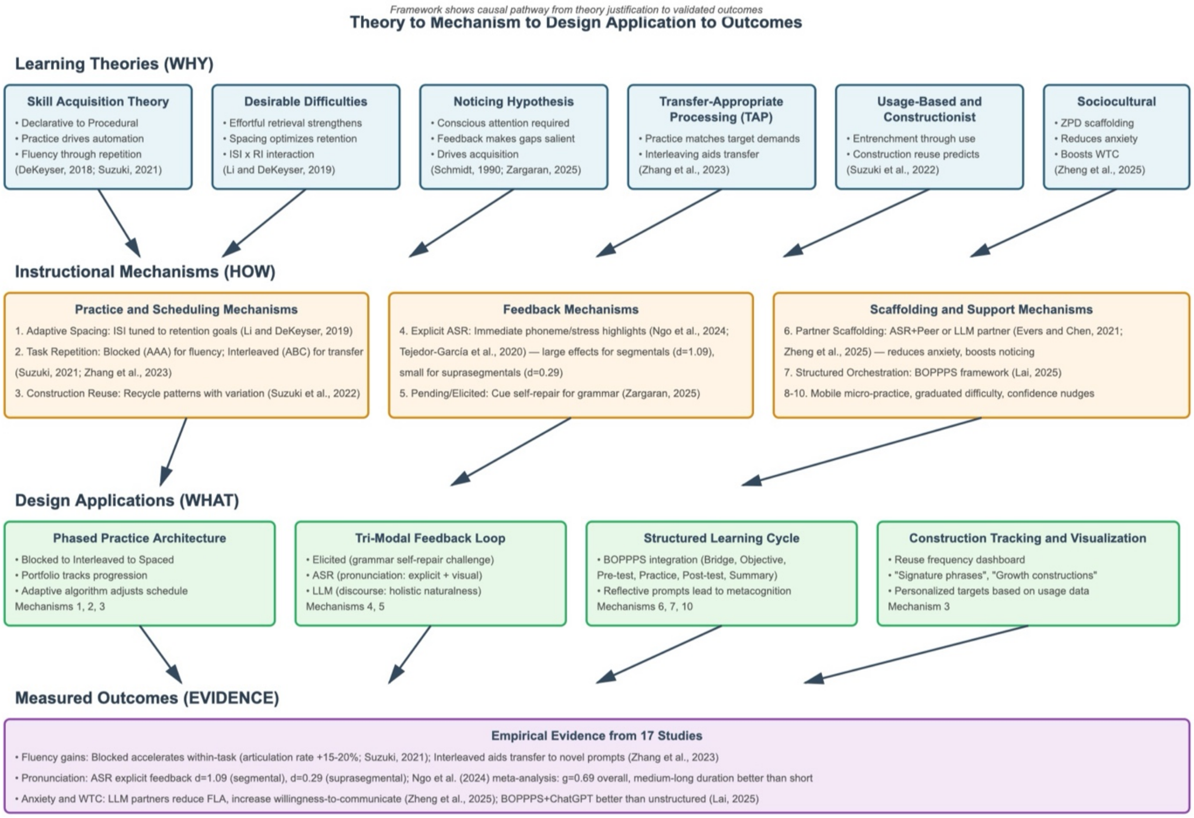
# 6. Theory-Mechanism Relationships

This section explicitly connects the theoretical justifications (WHY) with the proposed implementation mechanisms (HOW).

**Table 3:** Mapping of Learning Theories to Supporting Mechanisms

| Learning Theory (WHY) | Proposed Supporting Mechanisms (HOW) |
|---|---|
| **Skill Acquisition Theory** | **Task Repetition**: To provide the high-volume practice needed for proceduralization. |
| | **Practice Scheduling (Blocked)**: To accelerate automaticity for a specific, known task (Suzuki, 2021). |
| | **AI-Generated Feedback**: To offer continuous, real-time correction for skill refinement (Tejedor-García et al., 2020; Ngo et al., 2024; Evers & Chen, 2021). |
| **Retrieval Practice & Desirable Difficulties** | **Adaptive Spacing & Interleaved Scheduling**: To create effortful retrieval conditions that strengthen long-term memory and transfer (Li & DeKeyser, 2019; Yang et al., 2025). |
| **Usage-Based & Constructionist Theories** | **Task Repetition & Construction Reuse**: To facilitate the entrenchment of linguistic constructions through repeated activation (Suzuki et al., 2022). |
| **Noticing Hypothesis** | **Explicit ASR Feedback & LLM Feedback**: To make the "gap" between the learner's output and the target form explicit and salient (Ngo et al., 2024). |
| | **Pending Feedback**: To compel the learner to actively search for and consciously notice the correct form (Zargaran, 2025). |
| **Sociocultural & Affective-Motivational Theories** | **Partner Scaffolding (LLM/Peer):** To provide a non-threatening MKO that reduces anxiety and boosts confidence, WTC, and self-efficacy (Zheng et al., 2025; Evers & Chen, 2021). |
| | **Structured Orchestration (BOPPPS):** To create a predictable, supportive learning environment that may lower anxiety (Lai, 2025). |
| | **Mobile Micro-Practice**: To boost words-per-minute alongside band scores (Farooqi, 2025). |
| **Transfer-Appropriate Processing (TAP)** | **Interleaved Scheduling**: To practice switching between different contexts, which better resembles real-world communication and improves transfer to new tasks (Zhang et al., 2023). |

**Figure 2**: Theory-Mechanism-Design-Outcome Flow




*This figure provides a comprehensive visual synthesis of how the six learning theories identified in Section 4 connect to the ten instructional mechanisms described in Section 5, which then inform the four design applications outlined in Section 8, ultimately leading to measurable outcomes in pronunciation, fluency, discourse competence, and affective dimensions. The flow diagram makes explicit the logical chain from theoretical foundation through mechanism selection to practical implementation and expected outcomes.*

## 6.1 Notable Integration Points

- **Skill Acquisition Theory** supports a blocked-first, interleaved-later schedule (§5.1.2) with shorter ISI (§5.1.1) to optimize proceduralization before demanding transfer.

- **Sociocultural/Affective Theory** underpins both LLM partners (§5.3.1) and confidence nudges (§5.3.5), emphasizing the critical role of reducing anxiety and providing supportive feedback.

- **Transfer-Appropriate Processing** links interleaving (§5.1.2) with structured lesson scaffolds (§5.3.2) to ensure practice conditions match target performance demands.

- **Noticing Hypothesis** connects explicit ASR feedback (§5.2.1) with pending feedback (§5.2.2), showing that both explicitness and self-discovery enhance awareness of linguistic gaps.

See Appendix D for the study-by-mechanism matrix.

---

## 7. Key Findings

### Finding 1: Practice Scheduling as Primary Driver

The reviewed literature consistently demonstrates that the way practice is scheduled is as important as the practice itself. An optimal architecture begins with blocked practice to build foundational fluency and then transitions to interleaved practice to ensure skills are flexible and transferable. This progression requires a portfolio-based learning approach to track longitudinal progress and inform the transition between schedule types.

### Finding 2: Multi-Modal Feedback System for Holistic Competence

The evidence strongly suggests that a combination of feedback modalities is necessary to develop fully communicatively competent speakers:

- **ASR for Pronunciation:** Explicit, corrective ASR feedback is highly effective for grammatical competence of pronunciation (Ngo et al., 2024).
- **LLMs for Discourse:** LLMs can provide feedback on higher-level aspects of communicative competence, including discourse management, strategic language use, and sociolinguistic appropriateness (naturalness).
- **Pending/Elicited Feedback for Self-Regulation:** Prompting learners to self-correct is a direct application of metacognitive training, moving the learner from passive recipient to active participant in their own learning process (Zargaran, 2025).

### Finding 3: Structured Frameworks Transform AI into Coaching Partners

The research by Lai (2025) provides a critical insight: the effectiveness of an AI partner like ChatGPT is dramatically amplified when integrated into a structured pedagogical framework like BOPPPS. The study found that learners in the structured group showed significantly greater gains across all five speaking sub-skills compared to a group with unstructured AI access. The structure helps learners engage with the AI as a metacognitive dialogue partner, guiding them through intentional cycles of goal-setting, practice, and reflection.

### Finding 4: Construction Reuse Tracking Enables Precision Feedback

The work of Suzuki, Eguchi & de Jong (2022) identifies a powerful mechanism for fluency development. Their research shows that the repeated use, or reuse, of linguistic "constructions"—form-meaning pairings ranging from simple phrases ("I think...") to abstract grammatical patterns (POS trigrams)—is a direct predictor of fluency gains. By tracking which constructions a user practices and reuses, a system can move beyond simple metrics like "words per minute" to measure and guide the deep proceduralization of language itself.

## 8. Design Applications

While individual mechanisms (like ASR or practice scheduling) can be copied, a system built on a cohesive pedagogical philosophy is much harder to replicate. **The core differentiator is a design that shifts the focus from the AI giving feedback to the learner developing skills to self-assess.** This is achieved by building a system that intentionally trains metacognition (the ability to self-evaluate), develops holistic communicative competence (not just grammar), and uses a portfolio-based approach to make long-term growth visible and motivating.

### Application 1: Phased Practice Architecture

**Feature:** The system guides users from initial, blocked practice on a new topic to more challenging, interleaved practice over time. It then uses spaced repetition to ensure long-term retention.

**Evidence & Justification:** Supported by research on Skill Acquisition and Desirable Difficulties (Suzuki, 2021; Zhang et al., 2023). The ability to track progress across these phases is enabled by a portfolio-based learning approach.

**Proposed KPIs (Learning Outcomes)**

- +10% articulation rate and -15% mid-clause pauses by Week 4/6 on matched prompts.

- Maintain ≥75% of trained-task fluency metrics on unseen prompts (transfer check).

**Proposed KPIs (Product / Usage)**

- ≥70% of active users complete ≥3 sessions/week; average ISI 1–3 days.

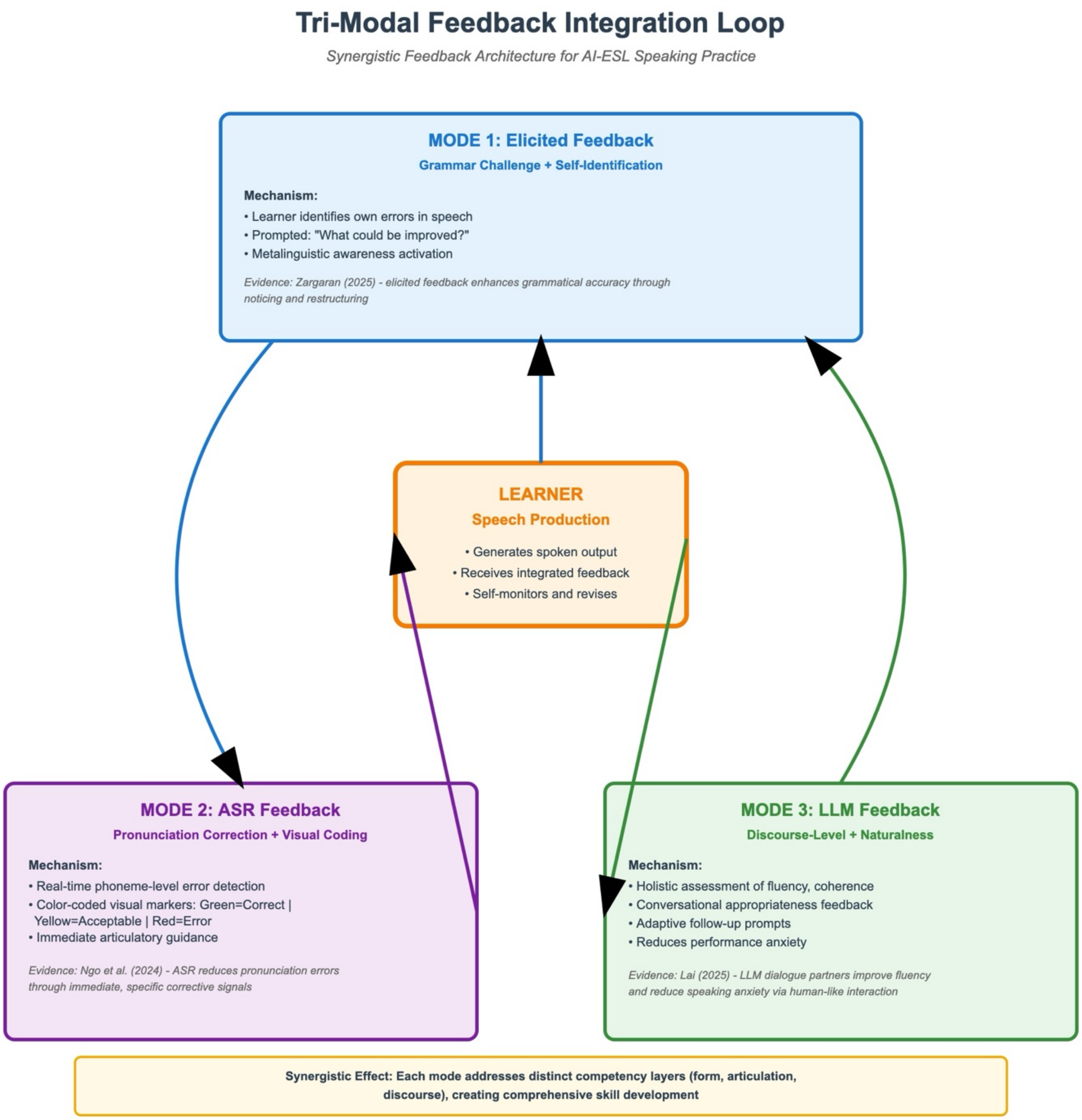
- ≥70% progress from blocked to interleaved mode by Week 2–3 without drop in fluency.

### Application 2: Tri-Modal Reflective Feedback Loop

**Feature:** A feedback loop that combines ASR, LLM, and elicited feedback to target different competencies and foster self-regulation:

- **Elicited Feedback ("Grammar Challenge"):** Before showing corrections, the user is prompted to find and fix their own grammatical errors.
- **Pronunciation Feedback (ASR):** Explicit, color-coded feedback on segmental pronunciation errors.
- **Discourse Feedback (LLM):** Holistic feedback on clarity, naturalness, grammar, and vocabulary in context.

**Figure 3**: Tri-Modal Feedback Integration Loop




*This figure illustrates the synergistic integration of three feedback modalities—elicited (grammar challenge), ASR (pronunciation correction), and LLM (discourse-level naturalness)—each addressing distinct competency layers (form, articulation, discourse) to create comprehensive skill development through continuous learner-centered iteration.*

**Evidence & Justification:** This integrated loop is designed to build full communicative competence. The elicited feedback component is a direct application of metacognition and self-regulated learning, training the user to self-monitor and evaluate.

### Proposed KPIs (Learning Outcomes)

- Pronunciation: ≥0.4 SD gain or 20% drop in low-confidence ASR flags.

- Grammar: ≥30% self-repair rate on prompted errors.

- Discourse: +0.5 on coherence / discourse rubric.

**Proposed KPIs (Product / Usage)**

- ≥70% of sessions include at least two feedback channels (ASR + pending or LLM).

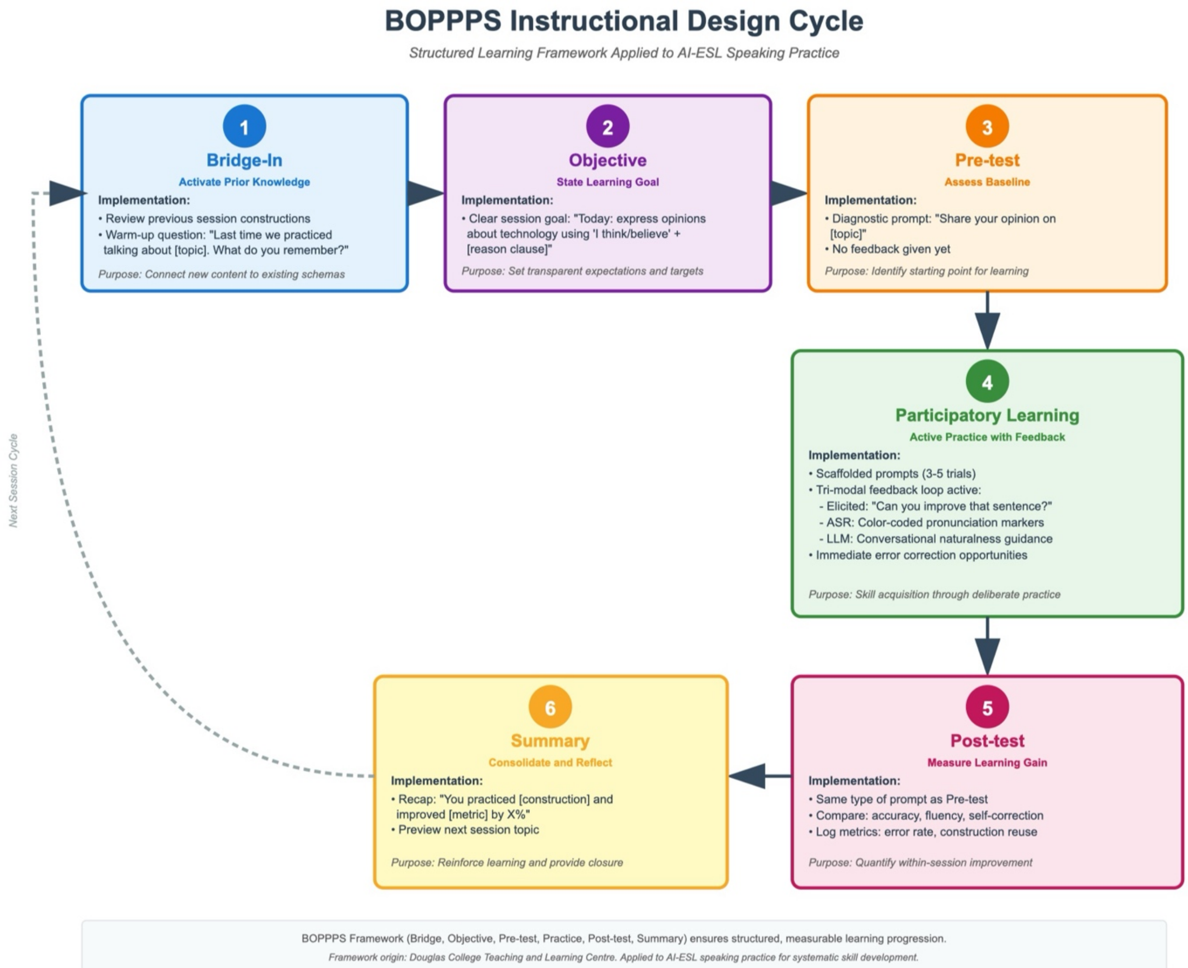
- ≥50% of sessions show at least one accepted/actioned suggestion; avg ≥2 re-records per session.

## Application 3: Structured Learning Cycle

**Feature:** Frame each practice session within a simplified BOPPPS-style learning cycle to promote intentional learning:

- **Bridge-In & Objective:** The user is presented with the topic and a clear goal (e.g., "Today's goal is to practice describing a past event using narrative tenses.")
- **Participatory Learning:** The user engages in the core speaking task and the Tri-Modal Feedback Loop.
- **Post-Assessment & Summary:** After reviewing feedback, the user is prompted with a single reflective question, such as "What is one pattern you noticed in your speech?" or "What will you focus on in your next session?" This response is saved as part of the session's portfolio entry.

**Figure 4**: BOPPPS Instructional Design Cycle




*This figure demonstrates the six-stage BOPPPS framework (Bridge-in, Objective, Pre-test, Participatory Practice, Post-test, Summary) applied to AI-ESL speaking practice. The visual representation shows how each stage creates structured, measurable learning progression with explicit feedback mechanisms at each phase, ensuring systematic skill development from activation of prior knowledge through consolidation and reflection.*

**Evidence & Justification:** Directly applies the findings of Lai (2025) to ensure the AI is used as a metacognitive partner rather than a passive tool. This structured reflection is a core component of building learner autonomy.

**Proposed KPIs (Learning Outcomes)**

- ≥60% of structured sessions show goal-aligned pre/post improvement in-session.

- Anxiety ↓0.5 and self-efficacy ↑0.5 on 5-point scales over the intervention.

**Proposed KPIs (Product / Usage)**

- ≥80% of structured sessions reach the reflection step.

- Time balance: ≥40% in speaking, ≥20% in feedback/reflection per session.

### Application 4: Construction Tracking and Visualization

**Feature:** The system's backend will identify and track the user's reuse of key grammatical and lexical constructions over time. This data will be made visible to the user in a dashboard.

**"My Speaking Patterns" Dashboard:** A section that visualizes the user's habits, showing:

- "Signature Phrases": Constructions the user relies on most frequently
- "Growth Constructions": New, more complex constructions the user is beginning to incorporate
- "Practice Targets": Useful constructions the user has encountered but not yet used

**Evidence & Justification:** Directly applies the findings of Suzuki et al. (2022). By making construction reuse—a key predictor of fluency—visible, the system gives users a concrete, actionable, and deeply personalized target for improvement.

**Proposed KPIs (Learning Outcomes)**

- 2× increase in target construction use by Session 4; +15% mean length of run.

- Increase in distinct constructions used correctly at least twice in spontaneous speech.

**Proposed KPIs (Product / Usage)**

- ≥60% of active users view the construction dashboard weekly; ≥40% launch practice from it.

- ≥30% of tracked constructions move up one status band over 8–12 weeks (weak → emerging → stable).

---

## 9. Discussion

### 9.1 Summary of Evidence

This scoping review synthesized 17 empirical studies, identifying six theoretical categories and ten evidence-based mechanisms. The evidence shows that practice architecture, not just volume, is critical. Blocked practice builds initial fluency, while interleaved practice promotes transfer. However, massed repetition can harm articulation speed (Li & DeKeyser, 2019; Suzuki & Hanzawa, 2022). Multi-modal feedback is also essential. ASR excels at segmental pronunciation (Ngo et al., 2024), while pending feedback and LLM dialogue can target grammar and discourse competence (Zargaran, 2025; Zheng et al., 2025). Finally, the effectiveness of AI tools is multiplied when they are embedded within structured pedagogical frameworks (Lai, 2025).

## 9.2 Evidence Quality and Weighting

Although all 17 included studies met the basic inclusion criteria, their methodological rigor and scope varied, and this variation informs how strongly specific conclusions should be interpreted. Several of the most influential findings in this review are supported by higher-quality evidence. For example, conclusions about **ASR-based pronunciation feedback** draw on a meta-analysis synthesizing 15 studies (Ngo et al., 2024) and multiple controlled interventions (Sun, 2023; Tejedor-García et al., 2020; Evers & Chen, 2021), which together provide robust support that explicit, automated feedback produces medium or larger improvements in segmental accuracy. Similarly, claims about **practice scheduling** and **distribution of practice** are grounded in carefully controlled experimental work by Suzuki and colleagues and by Li & DeKeyser (2019), including conceptual replications and delayed post-tests.

By contrast, some of the mechanisms most relevant for system design currently rest on a smaller number of promising but emerging studies. The **construction reuse** mechanism, for instance, is supported primarily by one mechanism-rich intervention (Suzuki, Eguchi, & de Jong), and the **pending feedback** mechanism relies on a single quasi-experimental study (Zargaran, 2025) conducted in an IELTS preparation context. These studies are rigorous in their own right but have not yet been replicated across diverse learner populations or platforms. As a result, design recommendations related to construction tracking and pending feedback should be treated as *high-potential hypotheses* that warrant empirical validation in future pilots rather than as universally established design laws.

## 9.3 Integration with Existing Literature

This review extends prior meta-analyses in three ways. First, it incorporates recent (2024-2025) LLM-based studies unavailable to earlier reviews. Second, it connects practice scheduling research from cognitive psychology to L2 speaking, demonstrating that principles like interleaving and spacing apply robustly (Zhang et al., 2023; Li & DeKeyser, 2019). Third, it uses a theory-mechanism-evidence framework to explain why interventions work, not just that they work, by tracing them back to foundational theories like Skill Acquisition and the Noticing Hypothesis.

## 9.4 Implications for AI-ESL System Design

The evidence supports three clear design imperatives for creating effective and defensible AI-ESL systems.

**Imperative 1: Architect Practice from Blocked to Interleaved.**

An effective system must be more than a random task generator. It should guide learners through a deliberate sequence, beginning with blocked practice to build foundational automaticity and transitioning to interleaved practice to ensure skills are flexible and transferable.

**Imperative 2: Integrate Feedback Across Multiple Modalities.**

No single feedback technology is sufficient. A robust system must function as an integrated feedback hub, leveraging ASR for pronunciation, pending prompts for grammatical self-repair, and LLM dialogue for discourse-level competence.

**Imperative 3: Amplify AI Effectiveness with Pedagogical Structure.**

The sophistication of an AI model is less important than the structure of the learning experience. Pedagogical orchestration is what transforms an AI from a passive tool into an active learning partner.

These implications point toward a defensible competitive strategy centered on integrated behavioral mechanisms requiring sustained engagement. Construction reuse tracking generates personalized insights only after users accumulate substantial practice history, creating switching costs. Adaptive scheduling gains precision through data accumulation, with parameter optimization improving as the system learns individual retention patterns.

## 9.5 Practical Considerations for Real-World Deployment

Translating the theories and mechanisms identified in this review into a deployable AI-ESL speaking system requires attention to infrastructure, cost, and data governance. The Real-Speaking prototype illustrates one feasible architecture that balances pedagogical ambition with technical and business constraints.

**ASR infrastructure.** Implementing explicit ASR feedback at scale requires choosing between local and cloud-based recognition. Local models (e.g., quantized Whisper running in the browser via WebGPU/WebAssembly) offer near-zero marginal cost and strong privacy, making it feasible to support the *volume* of practice required by Skill Acquisition Theory. However, their accuracy for heavily accented speech may lag behind state-of-the-art cloud APIs. Cloud ASR providers (e.g., commercial APIs similar to AssemblyAI or Soniox) offer higher accuracy and richer confidence measures but introduce per-minute costs. A hybrid approach—using local ASR for low-stakes, high-frequency feedback and cloud ASR for periodic “graded” assessments—can align costs with pedagogical priorities.

**LLM cost and scalability.** The tri-modal feedback loop relies on large language models to provide discourse-level feedback and conversational scaffolding. Fortunately, token-efficient models (e.g., lightweight instruction-tuned versions) make it possible to deliver several turns of dialogue and summary feedback for well under US$0.01 per 10–15-minute session, especially when prompts and context windows are carefully bounded. For institutional or commercial deployments, this cost structure is orders of magnitude lower than human tutoring and enables freemium or tiered pricing models, provided that usage is monitored and capped per user.

**Data privacy and speech recordings.** Spoken data are inherently sensitive. A local-first architecture, where raw audio and detailed transcripts remain on the learner's device by default (stored in secure browser storage or on-device databases), can lower privacy risk and anxiety. Only derived metrics—such as aggregated fluency scores, construction usage statistics, or anonymized error counts—need to be synced to the cloud for analytics and cross-device continuity. Clear consent flows, transparent explanations of where audio is stored, and simple data-export or deletion options are essential for building trust.

**Minimum viable dataset for construction tracking.** Usage-based mechanisms depend on repeated observations of learner language. In practice, meaningful construction tracking requires a minimum threshold of recorded speech—for example, on the order of several hundred utterances or 5–10 hours of practice per learner to reliably estimate which constructions are emerging or stabilizing. Early in a learner's journey, the dashboard should therefore emphasize simple indicators (e.g., total practice time, broad fluency trends), with more granular construction analytics activating only once sufficient data have been collected.

## 9.6 Conflicts and Tensions in the Evidence

The mechanisms synthesized in this review do not form a perfectly harmonious picture; several important tensions emerge across studies that have direct implications for system design. First, research on **massed versus distributed task repetition** reveals a trade-off between short-term fluency gains and long-term robustness. Massed or tightly blocked practice can produce the largest immediate reductions in pause length and increases in speed (Suzuki & Hanzawa, 2022), but these gains may degrade on delayed post-tests or lead to more verbatim repetition of practiced materials. More widely spaced or interleaved schedules support better transfer and retention but may slow early progress.

Second, comparisons of **blocked versus interleaved practice** indicate that no single schedule is universally superior. In some fluency-training studies, blocked practice outperforms interleaving during training and on near-transfer tasks, while in others, interleaved repetition leads to superior performance on new tasks and more flexible fluency (Zhang et al., 2023). These discrepancies suggest that the optimal schedule depends on the stage of learning, the type of outcome (speed vs. breakdown vs. repair fluency), and learners' tolerance for difficulty.

Third, the introduction of **AI-based feedback and dialogue partners** presents its own dualities. Automated feedback and AI partners often improve accuracy, fluency, and willingness to communicate, but they can also increase cognitive load if too many indicators, scores, or suggestions are presented simultaneously. Some learners may experience overload or frustration when confronted with highly granular feedback on every utterance. For designers, these tensions argue against fixed "one-size-fits-all" settings and in favor of adaptive systems that modulate practice intensity, mixing patterns, and feedback density based on observed learner response.

## 9.7 Limitations

Five categories of limitations constrain conclusions:

**1. Scope and Sampling**: The review was purposive, not exhaustive. It intentionally prioritized mechanism-rich studies to build a theoretical framework. A systematic follow-up review will be required to address comprehensive coverage and formal quality appraisal.

**2. Demographic Representation:** The evidence base is demographically narrow. Nine of the sixteen primary studies (56%) focused on East Asian university learners (Chinese, Japanese, and Taiwanese students), with ten studies (63%) conducted in East Asian contexts. This geographic and linguistic concentration limits the generalizability of the findings to learners from other linguistic and cultural backgrounds, particularly those from sub-Saharan Africa, Latin America, South Asia, and Eastern Europe (see Table 2 for detailed demographic breakdown).

3. **Cultural moderators**: Evidence is concentrated in East Asian university settings. Collectivist norms may make pending feedback and peer/AI dialogue feel safer, while individualist or low-power-distance contexts may prefer faster shifts to interleaving and more direct corrective feedback.

**4. Short Intervention Durations**: Most studies lasted fewer than twelve weeks, with none measuring retention beyond one month. This leaves long-term effects uncertain. Foundational research in cognitive psychology on memory consolidation suggests that true retention may only become apparent after 6 to 12 months.

**5. Methodological Variability:** The included studies vary widely in sample size, design rigor, and measurement tools. To provide transparency, a quality-appraisal rubric is included in Appendix B.

**6. AI-Assisted Review Process:** While the AI-assisted screening was reinforced with human triangulation, the process may have overlooked nuances. Full transparency regarding AI involvement is provided in the Acknowledgments.

## 9.8 Directions for Future Research

Four key areas for future research emerge from this review.

**1. Validate Findings with Diverse Populations:** Given that nine of the sixteen primary studies focused on East Asian university learners, the identified mechanisms should be validated with learners from underrepresented regions, including sub-Saharan Africa, Latin America, South Asia, and Eastern Europe, to assess cross-cultural validity and establish boundary conditions for different linguistic backgrounds.

**2. Conduct Longitudinal Studies:** Future work should include long-term follow-ups (6-12 months) to assess the retention of skills, which is crucial for validating the true impact of different practice schedules.

**3. Investigate Causal Mechanisms:** The link between construction reuse and fluency needs deeper investigation. Experimental designs could manipulate construction salience to better understand the causal pathway.

**4. Conduct Pilot Validation of Design Applications:** The design applications proposed in Section 8 require empirical validation. A single-group, pre-post feasibility study (N=15-20 users, 4-6 weeks) should be conducted to gather preliminary data on usability (SUS), engagement (session logs), and preliminary efficacy (pre-post fluency changes). This pilot would provide an empirical basis for a larger follow-on study.

### 9.9 Ethical and Bias Considerations

While this review focuses primarily on pedagogical mechanisms, deploying AI-powered speaking systems inevitably raises ethical questions. First, **accent bias in ASR** remains a persistent risk: acoustic models are typically trained on majority-accent corpora, which can lead to systematically lower recognition accuracy and harsher error flags for learners whose accents diverge from the training data. Designers should therefore monitor error patterns by L1 background, avoid over-interpreting raw ASR scores as direct measures of ability, and where possible incorporate human-calibrated scoring or accent-diverse training data.

Second, **cultural bias in LLM feedback** can shape what is presented as “natural” or “appropriate” speech. Large language models are trained on corpora that embed particular cultural norms and discourse styles; if left unchecked, they may implicitly penalize communication patterns that differ from these norms. For example, directness, hedging, or self-promotion may be evaluated differently across cultures. System prompts should explicitly instruct models to respect diverse registers and to highlight multiple acceptable options rather than a single “native-like” ideal.

Third, **privacy and ownership of speech data** are central to learner trust. Voice recordings and transcripts can reveal identity, emotional state, and sensitive personal information. A local-first architecture with clear consent mechanisms, minimal retention of raw audio on servers, and easy export/delete functionality helps align AI-ESL systems with emerging data-protection practices. Making these design choices visible to learners can also reduce anxiety and increase willingness to speak, thereby indirectly supporting the very mechanisms (e.g., increased practice volume and reduced affective filter) that this review highlights.

---

## 10. Conclusion

This scoping review demonstrates that effective AI-powered ESL speaking systems require integration of behavioral mechanisms grounded in cognitive and sociocultural learning theories rather than reliance on technological sophistication alone. The convergent evidence indicates that practice scheduling architecture (Zhang et al., 2023; Yang et al., 2025; Li & DeKeyser, 2019), multimodal feedback matching specific error types (Ngo et al., 2024; Sun, 2023; Zargaran, 2025), and structured pedagogical orchestration of AI interactions (Lai, 2025) produce measurable advantages in both immediate fluency and long-term transfer.

However, significant research gaps persist regarding optimal parameter settings for adaptive algorithms, causal mechanisms linking construction reuse to fluency development (Suzuki, 2017), and individual differences moderating intervention effectiveness. The field's rapid incorporation of LLM technologies (Zheng et al., 2025; Farooqi, 2025) presents both opportunity to address discourse-level competencies beyond pronunciation, and challenge to establish durable design principles that transcend specific model generations.

Future AI-ESL systems that embed evidence-based behavioral mechanisms while accumulating personalized learner data stand to create defensible competitive positions, but realizing this potential demands systematic research addressing identified gaps through adequately powered experimental trials with extended retention intervals and explicit attention to moderating factors.

---

## Acknowledgments

This review employed AI tools (GPT-5-Thinking, Grok-4, Gemini 2.5 Pro, Perplexity Comet) and CLI tools (Claude Code, Codex CLI, Gemini CLI) throughout the research process, adhering to principles of transparency and human oversight. AI assistance served four functions: (1) search task generation, (2) database search query refinement, (3) thematic categorization of retrieved records, and (4) full-text screening with structured scoring. All AI outputs were treated as draft proposals requiring human verification. All 57 search tasks, queries, filters, hit counts, and prompts were logged in structured CSV format (available as "Search_log_20251014.xlsx") to enable full replication.

For article selection, AI screening recommendations were validated through cross-model comparison and human spot-checking. Human oversight was critical during thematic synthesis—AI provided initial categorization, but the author made final determinations about theme boundaries, theoretical integration, and evidence-theory linkages. Key high-impact papers were read in full to extract insights that automated text processing might overlook.

During writing, raw ideas were drafted without focus on grammar, then AI tools proofread and improved clarity. The interactive process aligned outputs with the "Guide to Produce Scoping Reviews Using AI Tools."

---

## References


Abdelhalim, S. M., & Alsehibany, R. A. (2025). Integrating AI-powered tools in EFL pronunciation instruction: Effects on accuracy and L2 motivation. ***Computer Assisted Language Learning*.** https://doi.org/10.1080/09588221.2025.2534015

Evers, K., & Chen, S. (2021). Effects of automatic speech recognition software on pronunciation for adults with different learning styles. ***Journal of Educational Computing Research***, *59*(4), 669–685. https://doi.org/10.1177/0735633120972011

Farooqi, S.u.H. (2025). Efficacy of AI-generated feedback by SmallTalk2Me for improving speaking skill of Saudi EFL learners. ***Forum for Linguistic Studies***, *7*(3), 714–728. https://doi.org/10.30564/fls.v7i3.8294

Lai, Z. C.-C. (2025). Enhancing EFL oral proficiency through a ChatGPT-integrated BOPPPS learning framework. ***International Journal of Online Pedagogy and Course Design***. https://doi.org/10.4018/IJOPCD.383301

Li, M., & DeKeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 Mandarin tonal word production. ***The Modern Language Journa**l*, *103*(3), 607–628. https://doi.org/10.1111/modl.12580

Mingyan, M., Noordin, N., & Razali, A. B. (2025). Improving EFL speaking performance with an AI-powered mobile app in after-class assignments. ***Humanities and Social Sciences Communications***, *12*, 370. https://doi.org/10.1057/s41599-025-04688-0

Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. ***ReCALL***, *36*(1), 4–21. https://doi.org/10.1017/S0958344023000113

Schmidt, R. (1990). The role of consciousness in second language learning. ***Applied Linguistics***, *11*(2), 129–158.

Sun, W. (2023). The impact of automatic speech recognition technology on L2 pronunciation and speaking skills of EFL learners: A mixed methods investigation. ***Frontiers in Psychology***, *14*, 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. ***Language Learning***, *67*(3), 512–545. https://doi.org/10.1111/lang.12236

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. ***Language Learning***, *71*(2), 285–325. https://doi.org/10.1111/lang.12433

Suzuki, Y., & Hanzawa, K. (2022). Massed task repetition is a double-edged sword for fluency development: An EFL classroom study. ***Studies in Second Language Acquisition***, *44*, 536–561. https://doi.org/10.1017/S0272263121000358

Suzuki, Y., Eguchi, M., & de Jong, N. (2022). Does the reuse of constructions promote fluency development in task repetition? ***TESOL Quarterly***, *56*(4), 1290–1315. https://doi.org/10.1002/tesq.3103

Tejedor-García, C., Escudero-Mancebo, D., González-Ferreras, C., Cámara-Arenas, E., & Cardeñoso-Payo, V. (2020). Assessing pronunciation improvement in students of English using a controlled CAPT tool. ***IEEE Transactions on Learning Technologies***, *13*(2), 269–280. https://doi.org/10.1109/TLT.2020.2980261

Vygotsky, L. S. (1978). ***Mind in society***. Harvard University Press.

Yang, G., Wang, Y., Zhang, Y., Yang, M., Zeng, Q., & Song, Z. (2025). AI-supported interleaved training strategy for EFL impromptu speaking: Effects on performance, engagement, and acceptance. ***The Asia-Pacific Education Researcher***, *34*, 1519–1540. https://doi.org/10.1007/s40299-024-00962-y

Zargaran, Z. (2025). The effect of pending feedback on IELTS speaking: Grammatical range and accuracy. ***Australian Journal of Applied Linguistics***, *8*(1), 102484. https://doi.org/10.29140/ajal.v8n1.102484

Zhang, M., Yi, N., & Zhou, D. (2023). The effects of task repetition schedules on L2 fluency enhancement. ***Languages***, *8*(4), 252. https://doi.org/10.3390/languages8040252

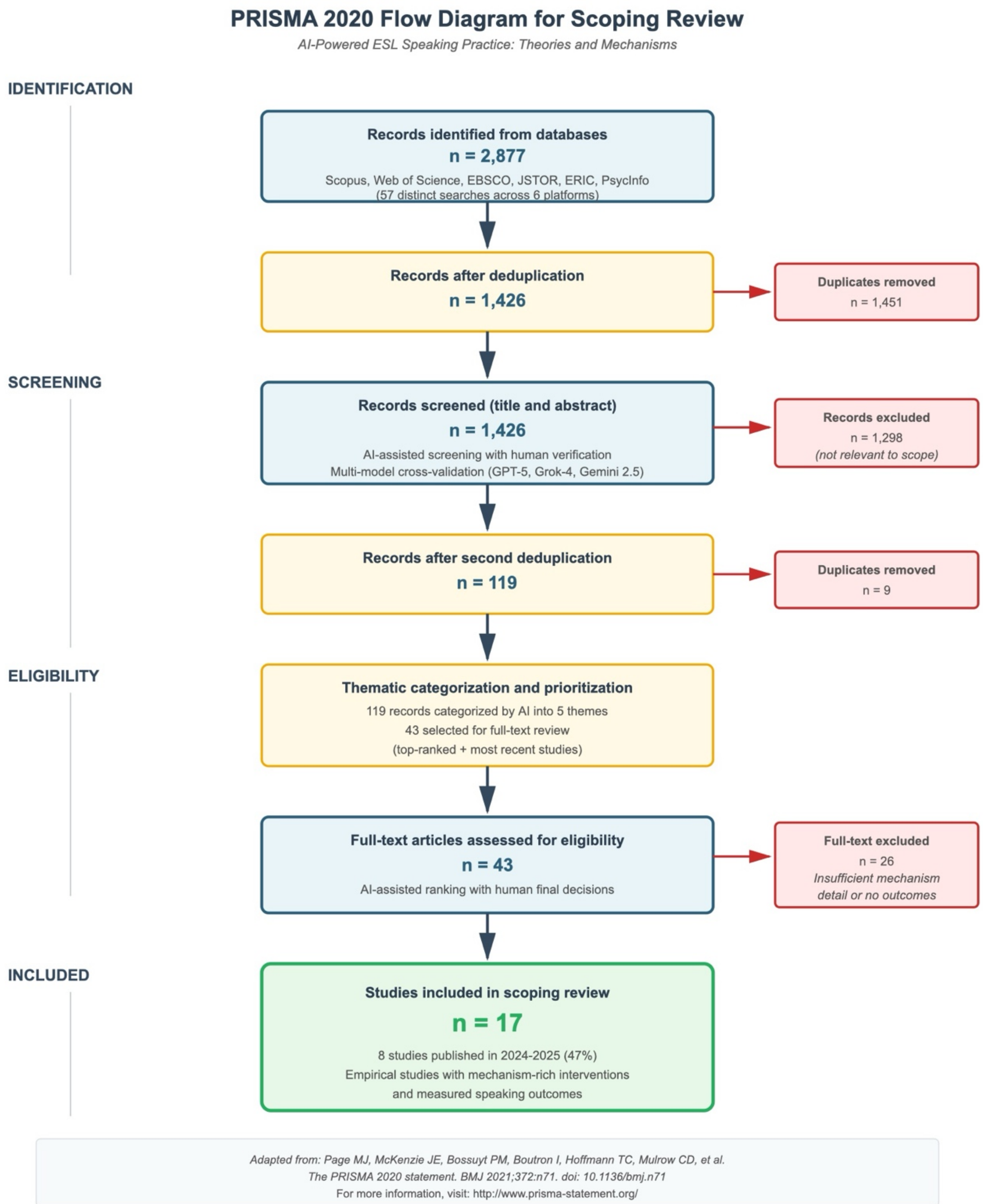
Zheng, Y.-B., Zhou, Y.-X., Chen, X.-D., & Ye, X.-D. (2025). The influence of large language models as collaborative dialogue partners on EFL oral proficiency and foreign language anxiety. ***Computer Assisted Language Learning.*** https://doi.org/10.1080/09588221.2025.2453191

---

# Appendix

## A.1 PRISMA Flow Diagram

**Figure 5**: PRISMA 2020 Flow Diagram for Scoping Review




*This diagram illustrates the complete study selection process following PRISMA 2020 guidelines. Starting with 2,877 records identified from six databases (Scopus, Web of Science, EBSCO, JSTOR, ERIC, PsycInfo), the review employed AI-assisted screening with multi-model cross-validation (GPT-5, Grok-4, Gemini 2.5 Pro) and human verification at all decision points. After initial deduplication (1,426 records), title/abstract screening (128 records), second deduplication (119 records), thematic prioritization (43 records selected for full-text review), and eligibility assessment, 17 studies were included in the final scoping review. Notably, 8 studies (47%) were published in 2024-2025, reflecting the recent acceleration in AI language learning research.*

## A.2 Complete Search Log

Complete documentation of all 57 search queries is available in the supplementary file "Search_log_20251014.xlsx"

# Appendix B Detailed Search Protocol and AI–Human Triangulation

This appendix documents the full search strategy, AI–human hybrid screening protocol, and PRISMA selection process that underlie the scoping review. It includes (a) database-specific search strings and filters, (b) chronological search logs and hit counts for each platform, (c) the three-layer AI–human triangulation workflow (model prompts, voting rules, and adjudication guidelines), and (d) the final PRISMA flow diagram with counts at each stage. Together, these materials provide an auditable trail that allows other researchers to replicate, critique, or extend the search in future work.

- **Databases Searched:** 57 distinct searches across 6 platforms: Scopus, Web of Science, Education Source (EBSCO), JSTOR, ERIC, and APA PsycInfo
- **Date of Last Search:** September 15, 2025
- **Search Strategy:** Targeted searches across three objectives:
    - Behavioral science interventions for skill acquisition and adherence
    - Practice scheduling and automaticity mechanisms
    - AI applications in ESL speaking practice

**Key Search Terms:** Combinations of: proceduralization, automaticity, speaking, fluency, second language; speech recognition, ASR, feedback; chatbot, conversational agent, LLM, ChatGPT; pronunciation, CAPT, assessment; implementation intention, self-regulated learning, skill acquisition, transfer of training

**Example Core Search Strings:**

- Scopus (SCOPUS-05): `TITLE-ABS-KEY(("speech recognition" OR ASR) AND (feedback) AND ("second language" OR ESL))` [2015–2025; English] → 69 hits, 69 exported

- Scopus (SCOPUS-06): `TITLE-ABS-KEY((pronunciation OR prosody) AND (feedback) AND (AI OR "intelligent tutor"))` [2015–2025; English] → 91 hits, 91 exported

- Web of Science (WOS-05): `(ESL OR EFL) AND speaking AND (AI OR "LLM" OR chatbot OR "artificial intelligence" OR "language model")` [2015–2025; English; Article/Proceeding] → 88 hits, 88 exported

- ERIC (ERIC-08): `(ESL OR EFL) AND (AI OR chatbot OR "conversational agent") AND (speaking OR "oral proficiency") pubyearmin:2015 pubyearmax:2025` [Peer-reviewed; Full-text on ERIC] → 12 hits, 12 exported

Complete search strings are available in "Search_log_20251014.xlsx."

**Filters Applied:** 2015-2025 (pre-2015 only for foundational mechanism sources when directly applied to speaking design decisions); English language; adult/higher education learners; peer-reviewed publications

**Total Records Identified:** 2,877 (across all 57 searches)

**Initial Retrieval:**

- Database hits: 2,877 records
- Exported count: 1,827 records
- After de-duplication: 1,426 unique records

**Title/Abstract Screening (AI-assisted with human verification):**

- The AI-assisted review process was reinforced by a human validation protocol to ensure credibility. The process involved:
    1. **Triangulation**: Key prompts for screening and thematic analysis were run across multiple AI models (including GPT-5 Thinking, Gemini 2.5 Pro, and Grok 4) to compare outputs and identify model-specific biases.
    2. **Human Oversight and Adjudication**: A human reviewer verified all AI-generated inclusion/exclusion decisions. For records flagged as "maybe," a second human reviewer resolved discrepancies. While formal inter-rater agreement was not calculated for this scoping phase, this multi-layered verification process reinforces the reliability of the AI-human collaboration.
- Initial screening round: 1,426 records → 110 includes, 1,277 excludes, 39 maybes
- Maybe resolution (3 rounds): 39 records → 11 additional includes, 21 excludes, 7 unresolved maybes carried forward
- Total after title/abstract screening: 128 records (121 definite includes + 7 maybes)

**Deduplication (post-screening):** 119 unique records after removing overlaps between databases

**Thematic Analysis (multi-AI comparison with human synthesis):**

- Categorized the 119 records by themes using AI analysis to prioritize studies for full-text review
- 43 records selected for full-text review focusing on five themes: 1. Generative AI and dynamic assessment for speaking 2. Practice scheduling and repetition for automaticity 3. ASR-enabled pronunciation for intelligibility 4. Task-based speaking with collaborative feedback 5. Feedback timing and knowledge type

**Full-Text Screening (AI-assisted review with human final decisions):**

- 43 records assessed for eligibility with AI rankings
- Top 11 ranking studies + 6 most recent studies were selected
- **Final Included Studies:** 17 studies

**Publication Recency:** Of the final 17 papers, 8 were published in 2024-2025, reflecting the recent surge in AI language learning research

**Inclusion Criteria:**

- Adult or university-level EFL/ESL learners
- AI-supported interventions (ASR, LLM, mobile apps) or theoretically grounded practice scheduling studies
- Measured speaking outcomes (pronunciation accuracy, oral fluency, discourse complexity, or affective measures)
- Empirical studies (experimental, quasi-experimental, mixed-methods)
- Published 2015-2025

**Exclusion Criteria:**

- Children or primary/secondary school learners as sole population
- Non-AI interventions without theoretical relevance to system design
- No measured speaking outcomes
- Reviews, commentaries, or purely theoretical papers
- Non-English publications

## Appendix C Quality Appraisal Matrix

This preliminary quality appraisal was developed post-hoc to document methodological variability across included studies. Future systematic reviews should incorporate formal quality appraisal at the study selection stage using established instruments (e.g., CASP, MMAT).

### Table C1: Quality Appraisal of Included Studies

| Study | Sample Size | Control Group | Measurement Validity | Attrition | Statistical Rigor | Overall Quality Rating |
|---|---|---|---|---|---|---|
| Li & DeKeyser (2019) | 80 (adequate) | Factorial design (high) | Validated tonal production measures (high) | Not reported (moderate) | Advanced mixed models (high) | **High** |
| Suzuki (2017) | 60 (adequate) | Yes, multiple conditions (high) | Miniature language paradigm with objective scoring (high) | Low reported attrition (high) | Appropriate inferential tests (adequate) | **High** |
| Suzuki (2021) | 68 (adequate) | Yes, control group (high) | Fluency metrics validated (high) | Not reported (moderate) | ANOVA with effect sizes (adequate) | **High** |
| Suzuki & Hanzawa (2022) | 79 (adequate) | Yes, multiple schedules + control (high) | Validated fluency measures (high) | Classroom setting, minimal attrition (high) | Appropriate tests with effect sizes (adequate) | **High** |
| Suzuki, Eguchi, & de Jong (2022) | 50 (adequate) | Comparative design (adequate) | Construction reuse coding validated (high) | Not reported (moderate) | Correlation and regression analyses (adequate) | **Adequate** |
| Zhang, Yi, & Zhou (2023) | 90 (adequate) | Yes, control group (high) | Fluency metrics (adequate) | Not reported (moderate) | ANOVA with post-hoc tests (adequate) | **Adequate** |
| Tejedor-García et al. (2020) | 18 (low) | Yes, classroom comparison (adequate) | CAPT automatic + human inter-rater reliability reported (high) | Not reported (moderate) | Appropriate for small n (moderate) | **Moderate** |
| Ngo, Chen, | 15 studies (meta- | Meta-analysis | Multiple outcome measures | Publication bias assessed | Meta-analytic procedures with | **High** |

| Study | Sample Size | Control Group | Measurement Validity | Attrition | Statistical Rigor | Overall Quality Rating |
|---|---|---|---|---|---|---|
| & Lai (2024) | analysis) | design (high) | aggregated (high) | (adequate) | moderator analysis (high) | |
| Sun (2023) | 61 (adequate) | Yes, comparison groups (high) | Multiple validated measures (high) | Low attrition reported (high) | ANCOVA with covariates (high) | **High** |
| Evers & Chen (2021) | 92 (adequate) | Yes, three conditions (high) | Validated ASR + human rating (high) | Reported and analyzed (high) | ANCOVA with moderation analysis (high) | **High** |
| Mingyan et al. (2025) | 63 (adequate) | Yes, control group (adequate) | App-generated + human rating (adequate) | Not reported (moderate) | Independent samples t-tests (adequate) | **Adequate** |
| Farooqi (2025) | 44 (moderate) | Pre-post design, no control (low) | IELTS bands + WPM (adequate) | Not reported (moderate) | Descriptive statistics, limited inferential (low) | **Low** |
| Abdelhalim & Alsehibany (2025) | 71 (adequate) | Pre-post design, no control (low) | Pronunciation measures + motivation scales (adequate) | Not reported (moderate) | Paired t-tests (moderate) | **Moderate** |
| Lai (2025) | 89 (adequate) | Yes, comparison group (high) | Multiple speaking sub-skills assessed (high) | Low attrition (high) | ANCOVA with effect sizes (high) | **High** |
| Yang et al. (2025) | 52 (adequate) | Yes, control group (high) | Rubric-based speaking assessment (adequate) | Not reported (moderate) | Structural equation modeling (high) | **High** |

| Study | Sample Size | Control Group | Measurement Validity | Attrition | Statistical Rigor | Overall Quality Rating |
|---|---|---|---|---|---|---|
| Zargaran (2025) | 60 (adequate) | Yes, two conditions (adequate) | IELTS grammatical accuracy rubric (adequate) | Not reported (moderate) | Correlation and ANOVA (adequate) | **Adequate** |
| Zheng et al. (2025) | 83 (adequate) | Yes, control group (high) | Validated oral proficiency + anxiety scales (high) | Not reported (moderate) | ANCOVA with multiple DVs (high) | **High** |

## Quality Rating Definitions

- **High:** Rigorous methodology with adequate sample size, appropriate control condition, validated measurement instruments, documented attrition, and sophisticated statistical analysis
- **Adequate:** Sound methodology with minor limitations (e.g., unreported attrition, moderate sample size, adequate but not optimal measurement)
- **Moderate:** Acceptable methodology with notable limitations (e.g., small sample, single-group design, limited statistical analysis)
- **Low:** Significant methodological limitations that constrain confidence in findings (e.g., very small sample, no control group, weak measurement, inappropriate statistics)

## Quality Appraisal Summary

Of the 17 included studies, 9 received a “High” quality rating, 5 received “Adequate,” 2 received “Moderate,” and 1 received “Low.” The predominance of high-quality studies (53%) strengthens confidence in the synthesis findings. However, methodological heterogeneity remains substantial, particularly regarding attrition reporting (11 studies did not report attrition data) and control group design (3 studies used pre-post designs without comparison groups).

The quality appraisal reveals that studies with longer intervention durations tended to report higher quality methodology (e.g., Sun, 2023; Lai, 2025; Evers & Chen, 2021 all received “High” ratings). Conversely, shorter pilot studies with pragmatic designs showed lower methodological rigor but provided valuable preliminary evidence for emerging AI technologies.

Future research should prioritize: (1) systematic attrition documentation and analysis, (2) preregistration of analysis plans to reduce researcher degrees of freedom, (3) power analyses justifying sample sizes, and (4) replication studies targeting high-impact findings from single studies (e.g., construction reuse mechanisms, pending feedback effects).

## Appendix D Mechanism × Studies × Evidence Strength Matrix

This matrix summarizes how strongly each of the 17 included studies supports key instructional mechanisms that inform AI-powered ESL speaking system design. Rows list the studies; columns list mechanisms. Each cell indicates the level of empirical support:

- ✓ = strong direct evidence for the mechanism in this study
- ~ = mixed, indirect, or context-limited evidence
- X = mechanism not tested or no support reported

### Mechanism Definitions

**M1 – Blocked → Interleaved Task Repetition.**
Comparisons of blocked vs. interleaved task schedules for speaking (e.g., A-A-A vs. A-B-C), including phased progressions from blocked to interleaved practice.

**M2 – Adaptive Spacing of Practice Sessions.**
Variations in inter-session interval (ISI) or temporal distribution of practice (massed vs. short-spaced vs. long-spaced) that affect acquisition and retention of spoken forms.

**M3 – Construction Reuse Tracking.**
Analyses of how repeated use of specific lexical or grammatical constructions predicts gains in fluency and complexity, and/or interventions that explicitly promote reuse.

**M4 – Explicit ASR-Based Pronunciation Feedback.**
Use of automatic speech recognition (or CAPT tools with ASR) to provide immediate, explicit feedback on segmental or suprasegmental pronunciation with measurable learning gains.

**M5 – Pending / Elicited Feedback.**
Feedback designs that delay or withhold the full correction, prompting learners to attempt self-repair before seeing the correct form.

**M6 – AI / Peer Collaborative Dialogue Partner.**
Use of AI (e.g., LLM-based chat) or structured peer collaboration as an active speaking partner that provides interaction, scaffolding, or co-reflection beyond one-way feedback.

**M7 – Structured Pedagogical Framework.**
Embedding AI or digital tools inside an explicit lesson structure (e.g., BOPPPS; staged cycles of pre-assessment → practice → feedback → reflection) rather than ad-hoc use.

**M8 – Mobile Micro-Practice (After-Class).**
Short, frequent, mobile or out-of-class speaking practice sessions driven by AI-powered tools (e.g., homework apps, after-class assignments).

## Table D1: Study × Mechanism Matrix

| Study (Abbrev.) | M1 Blocked → Interleaved | M2 Adaptive Spacing | M3 Construction Reuse | M4 ASR Pron. Feedback | M5 Pending Feedback | M6 AI / Peer Partner | M7 Structured Framework | M8 Mobile Micro-Practice |
|---|---|---|---|---|---|---|---|---|
| Suzuki & Hanzawa (2022) – Massed Task Repetition | X | ✓ | X | X | X | X | X | X |
| Suzuki, Eguchi & de Jong – Construction Reuse | ✓ | X | ✓ | X | X | X | X | X |
| Suzuki – Optimizing Fluency Training (Blocked vs. Interleaved) | ~ | X | X | X | X | X | X | X |
| Suzuki – Optimal Distribution of Practice (Morphology) | X | ✓ | X | X | X | X | X | X |
| Li & DeKeyser (2019) – Mandarin Tones | X | ✓ | X | X | X | X | X | X |
| Zhang et al. (2023) – Task Repetition Schedules | ✓ | X | X | X | X | X | X | X |

| Study (Abbrev.) | M1 Blocked → Interleaved | M2 Adaptive Spacing | M3 Construction Reuse | M4 ASR Pron. Feedback | M5 Pending Feedback | M6 AI / Peer Partner | M7 Structured Framework | M8 Mobile Micro-Practice |
|---|---|---|---|---|---|---|---|---|
| Yang et al. (2025) – AI-Supported Interleaved Training | ~ | X | X | X | X | ✓ | ~ | X |
| Ngo et al. (2024) – ASR Meta-Analysis | X | X | X | ✓ | X | ~ | X | X |
| Sun (2023) – ASR with Peer Correction | X | X | X | ✓ | X | ~ | X | X |
| Tejedor-García et al. (2020) – CAPT Tool | X | X | X | ✓ | X | X | X | X |
| Evers & Chen (2021) – ASR & Learning Styles | X | X | X | ✓ | X | ~ | X | X |
| Abdelhalim & Alsehibany (2025) – AI Pronunciation Tools | X | X | X | ✓ | X | X | X | X |
| Mingyan et al. (2025) – AI Mobile App (After-Class) | X | X | X | ✓ | X | X | X | ✓ |
| Farooqi (2025) – SmallTalk2Me AI Feedback | X | X | X | ~ | X | ✓ | X | ~ |

| Study (Abbrev.) | M1 Blocked → Interleaved | M2 Adaptive Spacing | M3 Construction Reuse | M4 ASR Pron. Feedback | M5 Pending Feedback | M6 AI / Peer Partner | M7 Structured Framework | M8 Mobile Micro-Practice |
|---|---|---|---|---|---|---|---|---|
| Lai (2025) – ChatGPT + BOPPPS | X | X | X | X | X | ✓ | ✓ | X |
| Zargaran (2025) – Pending Feedback & IELTS | X | X | X | X | ✓ | X | X | X |
| Zheng et al. (2025) – LLM Dialogue Partner & Anxiety | X | X | X | X | X | ✓ | X | X |

## Interpretation

- **Practice scheduling mechanisms (M1–M3).**
    - Blocked vs. interleaved task repetition (M1) is supported by multiple empirical studies, but with *conflicting patterns*: some show blocked practice advantages for within-training fluency (Suzuki – Optimizing Fluency Training), while others show interleaving advantages for transfer and overall fluency (Zhang et al., Construction Reuse).
    - Adaptive spacing (M2) has consistently strong support from Li & DeKeyser (2019) and Suzuki's morphology work, with additional evidence from massed vs. spaced task repetition (Suzuki & Hanzawa, 2022).
    - Construction reuse (M3) has one high-quality, mechanism-focused study (Suzuki, Eguchi & de Jong) showing that reuse frequency predicts fluency gains, making it a high-impact but still emerging mechanism.
- **Feedback mechanisms (M4–M5).**
    - Explicit ASR-based feedback (M4) has strong converging evidence: one meta-analysis (Ngo et al., 2024) plus multiple controlled studies (Sun, Tejedor-García et al., Evers & Chen, Abdelhalim & Alsehibany, Mingyan et al.). Evidence remains robust across different learner profiles and settings.

    - Pending / elicited feedback (M5) currently rests on a single but rigorous study (Zargaran, 2025), which reports clear gains in grammatical range and accuracy; this mechanism is promising but still under-replicated.
- **Scaffolding and deployment mechanisms (M6–M8).**
    - AI / peer partners (M6) show growing support: ASR studies indicate that peer-collaborative conditions often outperform solo practice, and recent LLM-based studies (Zheng et al., Lai, Yang et al., Farooqi) link AI dialogue partners to improved proficiency and reduced anxiety.

    - Structured frameworks (M7) have strong evidence from Lai (2025), where a ChatGPT-integrated BOPPPS framework outperforms unstructured AI use; Yang et al. (2025) offers additional but more indirect support via an AI-supported interleaved training strategy.

    - Mobile micro-practice (M8) is strongly exemplified by Mingyan et al. (2025) and moderately by Farooqi (2025), suggesting that short, after-class AI-mediated practice sessions contribute meaningfully to speaking gains, though more controlled comparisons are needed.

## Tiered Mechanism Prioritization for Design

Drawing on both the matrix above and the study quality ratings, mechanisms can be grouped into practical priority tiers:

- **Tier 1 – High-priority mechanisms (strong and replicated evidence).**
    - Adaptive practice scheduling (M1–M2) and explicit ASR-based pronunciation feedback (M4). These are supported by multiple high-quality experimental and meta-analytic studies and should anchor the initial system architecture.
- **Tier 2 – Promising, mechanism-rich additions (moderate evidence).**
    - Construction reuse tracking (M3), pending / elicited feedback (M5), structured lesson orchestration (M7), and mobile micro-practice (M8). Evidence is positive but limited to one or a few mechanism-focused studies; these features should be implemented with analytics to build further evidence.
- **Tier 3 – To monitor and refine.**
    - More complex scheduling refinements and advanced mastery gating that go beyond the basic blocked-to-interleaved and spacing patterns. These ideas are conceptually grounded but currently under-specified in the empirical literature and should be treated as experimental within pilots.