International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Disaggregating Speaking Gains: A Three-Arm Randomised Longitudinal Trial of AI-Only, Teacher-Led and Blended Speaking Instruction with Iraqi University EFL Learners

Muhannd Ali Abdullah
Head of International Relations Section, Planning and Follow up- up Department, Scientific Research Commission Baghdad, Iraq,
Corresponding Author: Muhannd Ali Abdullah, **E-mail**: *muhanned.a.abdullah@src.edu.iq*

| ABSTRACT
Artificial-intelligence (AI) applications for second-language speaking practice are rapidly expanding in higher education, including Iraq and the Kurdistan Region, where local evidence is dominated by perception surveys and small single-group designs. Although meta-analyses of dialogue-based computer-assisted language learning report moderate positive effects on L2 speaking, three instructional questions remain unresolved: whether AI practice produces broad speaking development or gains concentrated in temporal fluency; whether gains persist after practice ceases; and whether AI-supported instruction differs from time-equated teacher-led instruction rather than comparison conditions with unequal speaking time. This registered report proposes a three-arm, stratified cluster-randomised, longitudinal mixed-methods trial comparing AI-only, teacher-led, and blended AI–teacher speaking instruction under equivalent instructional time. Speaking outcomes are disaggregated into utterance fluency, pronunciation, grammatical accuracy, syntactic complexity, lexical diversity, and interactional competence, with a delayed post-test ten weeks after the intervention. Approximately 260 first- and second-year undergraduate EFL learners at CEFR A2–B1 level, nested within fifteen intact class clusters across three universities in Iraq and the Kurdistan Region, will be allocated to three arms. All arms will receive identical task content and approximately three hours of instruction weekly for fourteen weeks. A four-task speaking battery will be administered at pre-test, mid-point, immediate post-test, and delayed post-test. Temporal fluency will be extracted acoustically; pronunciation, comprehensibility, and interactional competence will be double-rated by trained raters with many-facet Rasch adjustment; and accuracy, complexity, and lexical measures will be derived from verified transcripts. Speaking anxiety, self-efficacy, willingness to communicate, self-regulated learning, and engagement will be measured at each occasion. Word error rate will be computed for each learner and modelled as a moderator of instructional effects. Linear mixed-effects models with learner and cluster random effects will estimate condition-by-time effects. ANCOVA will provide a confirmatory check, multilevel path-analytic mediation will use bootstrap intervals, and pre-specified false-discovery-rate control will be applied across a restricted primary outcome set. The design will provide a causal, sub-skill-level account of what each instructional mode produces, the first durability evidence for AI-mediated speaking gains following cessation, and evidence on whether automated-feedback validity functions as an individual-difference variable with instructional consequences.

| KEYWORDS
artificial intelligence; second-language speaking; utterance fluency; blended instruction; retention; automatic speech recognition; registered report; Iraq

| ARTICLE INFORMATION
ACCEPTED: 01 July 2026 **PUBLISHED:** 28 August 2026 **DOI:** 10.32996/ijllt.2026.9.9.2

1. Introduction

Speaking is the skill that learners of English as a foreign language most often report as the hardest to develop and the most anxiety-provoking to perform, and it is also the skill that conventional classroom conditions are least able to support. In a class of forty students with three contact hours a week, the arithmetic of turn distribution is unforgiving: individual speaking time is measured in minutes per week, corrective feedback on speech is necessarily sparse and public, and opportunities for extended,

low-stakes practice are close to nonexistent. This is not a failure of teaching; it is a structural constraint that most tertiary English programmes in the region operate under.

Artificial-intelligence applications appear to dissolve that constraint. An automatic-speech-recognition pronunciation tutor gives a learner unlimited private repetitions with immediate segmental feedback. A large-language-model voice agent will sustain a conversation for as long as the learner wishes, at whatever level of formality, without impatience and without an audience. The appeal is obvious, and adoption in Iraqi and Kurdistani higher education has moved considerably faster than evaluation. Survey work at four universities in the Kurdistan Region reports that a large majority of students believe their fluency has improved through such applications, that conversation practice is the feature they value most, and — a detail rarely followed up — that a substantial minority want better support for their accent (Vadivel et al., 2025). Comparable regional work is dominated by attitude questionnaires and by single-group pre-test–post-test designs (Rahman & Brime, 2025; Rusmiyanto et al., 2023; Leba & Butar-Butar, 2024).

Internationally, the evidence has advanced further. Three meta-analyses of dialogue-based and chatbot-assisted language learning now report moderate positive pooled effects on second-language speaking outcomes (Bibauw et al., 2022; Zhang et al., 2023; Hou & Min, 2026), and systematic reviews converge on pronunciation and fluency as the aspects most frequently improved, with immediate personalised feedback and reduced-anxiety practice conditions identified as the presumed active ingredients (Sutrah et al., 2025). At the level of "does this work at all," the question is effectively closed.

What has not closed is the set of questions a programme director actually faces. Should limited budget go to platform licences or to additional teacher-led speaking hours? If both, in what proportion? Does what students gain survive the end of the subscription? And — the question the Kurdistan survey respondents raised without anyone testing it — does the system hear these particular learners accurately enough for its feedback to be worth acting on?

None of these can be answered from the existing record, for four connected reasons. Speaking is measured holistically, so the profile of any gain is invisible. Measurement stops at the post-test, so durability is unknown. Comparison groups are typically non-artificial-intelligence applications or unmatched conventional classes, so instructional mode is confounded with time on task. And artificial-intelligence feedback is treated as a uniform, valid treatment, when in fact its content depends entirely on how accurately the underlying recogniser transcribes a given learner's accented speech — a quantity that varies substantially across first-language backgrounds (Bae & Kang, 2026) and that has never been measured inside an intervention study.

This registered report proposes a design built to answer all four. Its distinctive features are time equating across three arms including a genuine teacher-led comparator; decomposition of the speaking construct into six measured dimensions rather than one; a delayed post-test ten weeks after cessation with monitored abstinence; and per-learner recognition accuracy modelled as a moderator of learning gain.

2. Literature review

2.1 Effects of artificial-intelligence-mediated practice on second-language speaking

Meta-analytic work provides the appropriate starting point. Bibauw et al. (2022) synthesised dialogue-based computer-assisted language learning across skills and reported positive but heterogeneous effects, with task design and interaction type accounting for much of the variation. Zhang et al. (2023) reached similar conclusions for chatbot-assisted learning specifically. Hou and Min (2026), using a three-level model that permits multiple effect sizes per study, reported a moderate pooled effect on second-language speaking with stronger outcomes for goal-oriented and multimodal designs — and identified, as priorities for future work, longitudinal measurement and attention to the limits of speech recognition.

Primary studies in this area point in a consistent direction, though their designs share similar limitations. Qiao and Zhao (2023) randomised ninety-three Chinese learners to Duolingo-based or conventional instruction and found larger gains in speaking and self-regulation for the experimental group. Fathi et al. (2024) reported gains in speaking and willingness to communicate following artificial-intelligence-mediated interaction. El Shazly (2021) found reduced speaking anxiety alongside performance gains. A recent randomised controlled trial in the listening domain found that gains from speech-recognition-based instruction held at a three-week follow-up (Humanities and Social Sciences Communications, 2025), evidence that such follow-up measurement is both feasible and worth pursuing.

Sutrah et al. (2025), reviewing twenty-two studies published between 2020 and 2025, offer the clearest account of the field's methodological limits: sample sizes ranging from twenty to ninety-three, intervention periods of four to fourteen weeks, control groups often absent, and designs confined to a single site and a single application.

2.2 What is measured, and why the choice is substantive

The dominant outcome measure in this literature is a holistic rating — an International English Language Testing System band, a rubric total, or a teacher-assigned speaking grade. This is defensible for programme evaluation and indefensible for theory building, because it makes qualitatively different developmental profiles indistinguishable.

Second-language speaking research outside the technology literature has long treated speaking as multidimensional. The complexity–accuracy–fluency framework (Housen & Kuiken, 2009) separates dimensions that develop at different rates and trade

off against one another under different task conditions. Within fluency, Segalowitz's (2010) distinction between cognitive fluency, utterance fluency and perceived fluency locates the measurable construct in the speech signal itself: speed measures such as speech rate and articulation rate, breakdown measures such as pause frequency, duration and location, and repair measures such as repetitions and self-corrections. These are extractable automatically and with high reliability (de Jong & Wempe, 2009; de Jong et al., 2021), which removes them entirely from the domain of rater subjectivity.

In pronunciation research, the distinction between comprehensibility (listener-judged ease of understanding), intelligibility (actual understanding) and accentedness (perceived difference from a reference variety) is well established and consequential (Derwing & Munro, 2005; Levis, 2018; Kang et al., 2020, 2025). Comprehensibility is the pedagogically meaningful target; accentedness is a descriptive property that carries no necessary implication for communicative success.

Interactional competence — turn-taking management, topic development, repair, listener response — is a further dimension that a monologic rubric cannot capture at all, and it is precisely the dimension on which human and machine interlocutors might be expected to differ.

The prediction that follows is straightforward and, to our knowledge, untested: an intervention that multiplies private retrieval practice with immediate form-focused feedback should move speed and breakdown fluency and segmental accuracy substantially, while leaving syntactic complexity and interactional competence comparatively untouched. A holistic score would register this as a modest general improvement and would conceal the mechanism entirely.

2.3 Affective outcomes and their measurement problem

Reduced anxiety is the most frequently proposed explanation for artificial-intelligence benefits: practice with a machine removes the social-evaluative audience that drives foreign-language speaking anxiety (Horwitz et al., 1986). Reported increases in willingness to communicate (MacIntyre et al., 1998) and self-regulation (Qiao & Zhao, 2023) point the same way.

The difficulty is that these constructs are almost always measured by self-report administered to learners who know which condition they are in and who can readily infer what the researcher expects. In the regional literature this problem is acute: the principal Kurdistan study infers fluency development itself from student self-report (Vadivel et al., 2025). Where the outcome and the mechanism are both self-reported by unblinded participants, the resulting correlation is uninterpretable. The present design therefore retains validated self-report scales but triangulates them with behavioural indices derived from the speech signal.

2.4 Blended designs and the comparator problem

Reviews converge on the recommendation that artificial intelligence be integrated within a blended ecosystem complementing human instruction rather than adopted as a standalone tool (Sutrah et al., 2025). The theoretical warrant is sociocultural: a machine supplies feedback but not mediation, and it is the teacher who converts machine output into a tool the learner can act on.

One recent study has tested a version of this claim directly. A sixteen-week quasi-experimental study with one hundred and fifty Chinese learners compared artificial intelligence with teacher scaffolding, artificial intelligence alone, and a non-artificial-intelligence gamified control, and reported that the scaffolded condition achieved greater and more sustained proficiency gains, with motivation mediated by satisfaction of self-determination needs (Frontiers in Psychology, 2025). This is an important precedent and we position the present study explicitly against it.

Three limitations of that design motivate the present one. Its comparison condition was a non-artificial-intelligence gamified platform rather than teacher-led communicative speaking instruction, so the study cannot speak to the substitution question institutions actually face. Its outcome was a holistic proficiency measure, so the profile of gains is unknown. And its measurement ended with the intervention, so durability is unaddressed.

2.5 Durability

Almost no study in this literature measures learners after practice stops. This matters theoretically as well as practically. Gains attributable to proceduralisation of retrieval should decay in the absence of continued practice; gains attributable to reduced anxiety or increased self-efficacy have no obvious reason to decay at the same rate. The two accounts therefore make different predictions about the shape of the post-cessation profile, and a delayed post-test discriminates between them. Longitudinal fluency research outside the technology literature shows that different fluency measures both emerge and stabilise on different

2.6 The unexamined assumption: feedback validity

Every mechanism proposed for artificial-intelligence benefit — noticing, modified output, increased perceived control — presupposes that the feedback the learner receives is accurate. That presupposition is not safe. Automatic speech recognition performs measurably worse on second-language-accented speech than on the varieties dominating its training data, and the degradation differs by first language: Bae and Kang (2026) found that transcription accuracy varied systematically by first-language background across two commercial systems, while human expert raters showed no comparable bias. Arabic-first-language phonological features have been shown to affect recognition accuracy in ways that neither teachers nor learners anticipate correctly (Speech Communication, 2024). Automated scoring research has long treated subgroup differences in

machine–human agreement as a fairness question requiring explicit investigation (Loukina & Buzick, 2017). timescales, which makes the disaggregated approach a prerequisite for asking the durability question at all.

The consequence for intervention research is direct. If recognition fails for a subset of learners, those learners receive degraded or incorrect corrective feedback while nominally receiving the same treatment as everyone else. Treatment fidelity, in other words, varies at the level of the individual learner in a way that the study design cannot see unless it measures recognition accuracy. We are aware of no intervention study that has done so.

2.7 The regional evidence base

Work in Iraq and the Kurdistan Region is growing quickly. Abdulhussein Dakhil et al. (2025) reported effects of artificial-intelligence-mediated speaking assessment on speaking performance and willingness to communicate with a randomised sample of forty learners. Rahman and Brime (2025) surveyed two hundred fourth-year students in three Kurdistani universities on their perceptions of artificial-intelligence tools. Vadivel et al. (2025) surveyed one hundred and five students across four universities. Aryanti and Santosa (2024) reviewed ten studies of pronunciation applications.

The pattern is one of rapid, useful mapping work with, as yet, no study combining a powered sample, random allocation, an active time-equated comparator, objective speech measurement and post-intervention follow-up. Given that the region's learners are simultaneously among the least represented in the training data of the systems being marketed to them, the absence of rigorous local evidence is not merely a regional gap but a fairness concern.

2.8 The present study

We therefore propose a study that holds instructional time constant across three arms, measures six speaking dimensions objectively, follows learners ten weeks past the end of the intervention, and measures the accuracy of the feedback each learner actually received. To our knowledge, no published study has combined these four features.

3. Theoretical framework

Four theories are retained; each generates a distinct, falsifiable prediction. Technology-acceptance models are deliberately excluded, since they predict adoption intention rather than learning gain and would add constructs without adding predictions.

3.1 Interaction, output and noticing. Long's interactionist account, Swain's output hypothesis and Schmidt's noticing hypothesis converge on the claim that pushed output plus feedback on that output drives development, and that repeated retrieval proceduralises access. Prediction: conditions maximising speaking turns and feedback density produce gains concentrated in speed and breakdown fluency and in segmental accuracy, and those gains, being practice-dependent, show partial decay once practice stops.

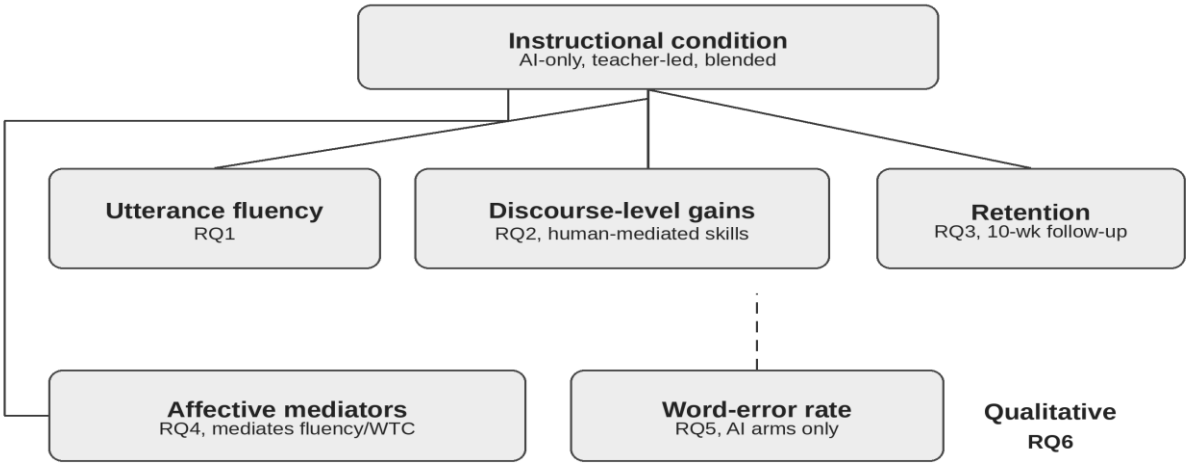
3.2 Sociocultural mediation. Vygotskian theory holds that development occurs through mediated activity within the zone of proximal development. An automated system supplies feedback but does not mediate: it does not calibrate across turns to what this learner can nearly do, and it does not negotiate meaning about pragmatic appropriacy. Prediction: the blended condition outperforms the artificial-intelligence-only condition specifically on comprehensibility and interactional competence, and shows the least decay, because the teacher converts machine output into a usable tool and builds strategies that persist.

3.3 Control-value theory and self-determination theory. Pekrun's control-value theory locates achievement emotions in appraisals of control and value; self-determination theory locates motivation in autonomy, competence and relatedness. Private practice raises perceived control and removes the social-evaluative threat central to speaking anxiety; learner control over topic and pace supports autonomy. Prediction: anxiety reduction and self-efficacy increase partially mediate the effect of artificial-intelligence-containing conditions on fluency and willingness to communicate, and affective change persists at follow-up.

3.4 Self-regulated learning. Zimmerman's cyclical model implies that practice converts to learning only when learners plan, monitor and act on feedback. Prediction: self-regulated learning functions both as an outcome (does the condition build it?) and as a moderator (does it condition the benefit?), with the artificial-intelligence-only arm most dependent on pre-existing self-regulation because no teacher supplies external regulation.

3.5 Feedback validity as a boundary condition. All four mechanisms take the same input: the feedback signal. If recognition is inaccurate, the noticing pathway is attenuated (the learner notices a gap that does not exist, or fails to notice one that does) and the control-value pathway may reverse (repeated misrecognition lowers perceived control and invites attribution of failure to oneself). Word error rate therefore enters the framework as a moderator on the path from condition to outcome, and as a predictor of feedback trust in the qualitative strand.

Figure 1 presents the integrated model. Each arrow corresponds to a numbered research question.



4. Research questions and hypotheses

RQ1. Do the three instructional conditions differ in gains on utterance-fluency indices from pre-test to immediate post-test? H1. Both artificial-intelligence-containing conditions will exceed teacher-led instruction on speech rate, mean length of run and reduction in silent-pause frequency, owing to greater individual speaking time. The blended and artificial-intelligence-only conditions will not differ substantially on these indices.

RQ2. Do the conditions differ on pronunciation, grammatical accuracy, syntactic complexity, lexical diversity and interactional competence? H2a. Artificial-intelligence-only will equal or exceed teacher-led instruction on segmental pronunciation accuracy. H2b. Blended will exceed both other conditions on comprehensibility and interactional competence. H2c. Conditions will not differ meaningfully on syntactic complexity over fourteen weeks.

RQ3. Which gains are retained ten weeks after practice ceases, and does retention differ by condition? H3. Blended will show the smallest decay overall; artificial-intelligence-only will retain temporal-fluency gains but show greater decay in accuracy and pronunciation; affective changes will decay least of all measured variables.

RQ4. Do speaking anxiety, self-efficacy and behavioural engagement mediate condition effects? H4. Anxiety reduction and self-efficacy increase will partially mediate effects of artificial-intelligence-containing conditions on fluency and willingness to communicate; log-derived engagement will mediate the artificial-intelligence-only effect more strongly than the blended effect.

RQ5. Does per-learner automatic-speech-recognition accuracy moderate the benefit derived from artificial-intelligence-containing conditions? H5a. Word error rate will be substantially higher for spontaneous than for read-aloud speech and will differ by first-language group. H5b. Learners with higher word error rates will show smaller pronunciation and accuracy gains in artificial-intelligence-containing conditions; no such moderation will appear in the teacher-led condition.

RQ6. How do learners and teachers account for the usability, trustworthiness and integration of artificial-intelligence feedback, and how do these accounts explain the quantitative patterns? Exploratory; no directional hypothesis.

No research question treats speaking proficiency as a single undifferentiated variable, and H5b is stated so that it can fail.

5. Method

5.1 Design

A three-arm, stratified cluster-randomised controlled trial with four measurement occasions and an embedded convergent mixed-methods component (quantitatively dominant; qualitative data collected at immediate and delayed post-test to explain quantitative patterns). Reporting will follow the CONSORT extension for cluster randomised trials, adapted for educational settings.

Randomisation is at the level of intact class clusters because within-class allocation cannot be implemented without contamination; the analysis accounts for clustering throughout (§6). The protocol will be pre-registered on the Open Science Framework prior to any data collection, with hypotheses, primary outcomes and the analysis plan time-stamped.

Table 1. Measurement schedule

Occasion	Timing	Speaking battery	Questionnaires	Other
T1 pre-test	Week 0	Full (4 tasks)	Full set	Placement test; digital-literacy and access survey.
T2 mid-test	Week 7	Abbreviated (monologue only)	Anxiety, self-efficacy	Platform logs (continuous)
T3 post-test	Week 14	Full, parallel form	Full set	Interviews; teacher focus groups; fidelity summary
T4 delayed	Week 24	Full, parallel form	Full set	Abstinence/usage log; follow-up interviews

5.2 Setting

English-language and English-for-academic-purposes programmes at three universities: two in the Kurdistan Region with predominantly Kurdish-first-language intake and one in central Iraq with predominantly Arabic-first-language intake. Multi-site recruitment serves two functions: it enables the first-language comparison required by RQ5, and it protects external validity against single-institution effects.

5.3 Participants

First- and second-year undergraduates, typical age eighteen to twenty-two, screened to the Common European Framework of Reference A2–B1 band, of mixed gender, device ownership and connectivity. Target subsamples of at least eighty Kurdish-first-language and eighty Arabic-first-language learners.

Inclusion criteria: enrolment in a required speaking or oral-communication course; placement score within the A2–B1 band; access to a smartphone or laptop with microphone; written informed consent.
Exclusion criteria: prior or current sustained use of an artificial-intelligence speaking tutor (more than two hours weekly for more than one month); residence exceeding three months in an English-dominant country; diagnosed speech or hearing impairment that would confound acoustic measurement (such students participate fully in classes; their data are excluded from acoustic analysis only, and this is reported); placement at C1 or above.

5.4 Sample size

Planning parameters, to be confirmed by simulation before recruitment:
•Target effect: condition-by-time interaction of f = 0.15, consistent with moderate pooled effects attenuated by an active comparator.
• α = .05, power = .90, three groups by four occasions, assumed within-subject correlation r = .60 → approximately 120 learners at the individual level.
• Design effect for cluster randomisation: DEFF = 1 + (m − 1) × ICC. With mean cluster size m = 18 and a conservative ICC = .05, DEFF ≈ 1.85 → approximately 222 learners.
• Attrition allowance of 15% to T3 and a further 10% to T4 → recruit approximately 260 learners in 15 clusters, five per arm.
The moderation test (H5b) is the least well-powered analysis; interaction effects require substantially larger samples than main effects. RQ5 moderation is therefore designated pre-specified but exploratory, reported with confidence intervals and accompanied by a sensitivity analysis stating the minimum detectable interaction effect size.
All figures in this subsection are planning assumptions, not results.

5.5 Randomisation, allocation concealment and blinding

Clusters are stratified by site, by baseline mean placement score (median split) and by majority first language, then randomly allocated within strata by an independent statistician using a computer-generated sequence with concealed allocation. Learners and teachers cannot be blinded. Raters, transcribers and acoustic analysts are blinded to condition and occasion: files are coded, shuffled and presented in random order. The analyst will specify the primary models before condition labels are unmasked.

5.6 Conditions

All three arms receive identical total instructional time (approximately three hours weekly for fourteen weeks, approximately forty-two hours), identical task content drawn from a common task bank, and identical assessment. Only the mode of practice and the source of feedback differ. Time equating is the design's central safeguard and is documented in the manuscript.

Table 2. Intervention conditions

	Arm A: AI-only	Arm B: Teacher-led	Arm C: Blended
Weekly structure	3 h supervised laboratory practice with AI tools; no teacher feedback on speech	3 h communicative speaking class: pair and group tasks, teacher corrective feedback, peer interaction	1.5 h AI laboratory practice + 1.5 h teacher-led session debriefing AI feedback
Teacher role	Attendance and technical support only; no language feedback	Full instructional role	Interprets AI reports, targets recurrent errors, runs interactional tasks the system cannot support
Feedback source	Automated (ASR and LLM)	Human	Both, with human mediation of machine output
Task content	Common bank	Common bank	Common bank
Total time	≈ 42 h	≈ 42 h	≈ 42 h

Arm B is not a deprivation condition: it is the current standard of instruction. All arms receive full access to the artificial-intelligence tools after T4.

5.7 Weekly cycle

Each week follows a common six-stage cycle so that content is held constant: (1) functional and topical focus introduced; (2) controlled practice (read-aloud and minimal pairs; automated in Arms A and C, teacher-led in Arm B); (3) semi-controlled monologue; (4) interactive task (Arm A: voice-agent role-play; Arm B: paired human role-play; Arm C: rehearsal with the agent followed by human performance); (5) feedback and repair episode; (6) brief reflective log entry, all arms, approximately five minutes.

5.8 Artificial-intelligence tools

A two-tool stack, version-locked and documented with retrieval dates:

1. An automatic-speech-recognition pronunciation tutor providing phoneme-level scoring, word-stress and intonation feedback, and adaptive drill sequencing, with exportable progress reports.
2. A large-language-model voice conversational agent supporting open and task-based dialogue with post-task feedback.

Where an application-programming interface is used, the system prompt is fixed and reproduced in Appendix B.

Model versions, configuration, prompt texts and any mid-study updates will be reported; unannounced vendor updates are declared as a limitation.

5.9 Teachers and standardisation

Where numbers permit, each teacher will teach in more than one arm, so that teacher identity is crossed with rather than nested within condition — the strongest available control for instructor effects. Where this is impossible, teacher is modelled as a random effect and the confound is declared. All teachers receive a six-hour training workshop and a manualised scheme of work; Arm C teachers receive additional training in reading and mediating automated feedback reports.

5.10 Speaking battery

Administered identically at each occasion, with counterbalanced parallel forms pilot-equated in advance:

1. Read-aloud passage (approximately 120 words) — supports segmental analysis and the read-speech recognition condition for RQ5.
2. Picture-based narrative monologue (one minute preparation, two minutes speech) — principal source of fluency and complexity–accuracy–fluency indices.
3. Unprepared opinion monologue (two minutes, no preparation) — low-planning condition; spontaneous-speech recognition condition for RQ5.
4. Paired interactive task (five minutes, matched interlocutor rotated across occasions) — sole source of interactional-competence data.

Recording equipment, microphone type, sampling rate and room conditions are standardised across sites and specified in the manuscript, since acoustic measures are sensitive to recording conditions.

5.11 Outcome measures

Table 3. Measurement instruments and extraction method

Construct	Measure	Extraction	Occasions
Speed fluency	Speech rate; articulation rate	Praat, automated syllable-nuclei detection (de Jong & Wempe, 2009; de Jong et al., 2021)	T1–T4
Breakdown fluency	Silent-pause frequency and mean duration (threshold 250 ms); phonation-time ratio; mean length of run; mid- vs end-clause pause ratio	Praat + clause-boundary coding	T1–T4
Repair fluency	Repetitions, self-corrections, false starts per 100 words	Transcript, double-coded	T1–T4
Pronunciation (objective)	Segmental substitution and deletion counts; word-stress errors	Transcript vs. phonemic target, read-aloud task	T1–T4
Pronunciation (rated)	Comprehensibility; accentedness (9-point scales, rated separately)	Trained raters, double-rated	T1–T4
Grammatical accuracy	Errors per 100 words; proportion error-free AS-units	Transcript coding	T1–T4
Syntactic complexity	Clauses per AS-unit; mean length of AS-unit (Foster et al., 2000)	Transcript segmentation, human-verified	T1–T4
Lexical diversity/sophistication	MTLD or MATTR; frequency-band coverage	Automated from verified transcripts	T1–T4
Interactional competence	Analytic scale: turn-taking, topic development, repair, listener response	Trained raters, double-rated, paired task	T1–T4
Speaking anxiety	Validated speaking-anxiety scale (Horwitz et al., 1986, speaking subscale or adaptation)	Self-report	T1–T4
Willingness to communicate	Validated L2 WTC scale (MacIntyre et al., 1998)	Self-report	T1–T4
Speaking self-efficacy	Validated scale	Self-report	T1–T4
Self-regulated learning	Speaking-focused SRL scale	Self-report	T1, T3, T4
Engagement	Sessions, minutes on task, turns, words, feedback items, retries	Platform logs	Continuous
Feedback validity	Word error rate per learner per task type	ASR output vs. verified gold transcript	T1, T3

Scores produced by the commercial platform are collected but are never used as study outcomes; they serve only as the object of the validation analysis in §6.5.

5.12 Rubric and rater procedures

An analytic rubric with separate scales for comprehensibility, accentedness, grammatical accuracy, lexical range, discourse coherence and interactional competence, anchored to Common European Framework descriptors and to local benchmark samples. Scale functioning will be examined by many-facet Rasch analysis on a pilot of twenty recordings before main rating (Linacre, 1994).

Four raters, trained to a benchmarked standard. All recordings double-rated where resources permit, with a minimum thirty per cent double-rated overlap otherwise. Reliability reported as ICC(2,k) for continuous ratings and quadratic-weighted kappa for ordinal judgements; fair-average scores adjusted for rater severity via many-facet Rasch modelling are used in all analyses. Transcription and complexity–accuracy–fluency coding double-coded on twenty per cent of files with agreement statistics reported.

5.13 Qualitative data

Semi-structured interviews at T3 with a maximum-variation subsample (approximately twenty-four to thirty learners: high and low gainers, crossed by arm and first language) and at T4 with approximately twelve learners regarding maintenance. Teacher focus groups (approximately eight to ten teachers) at T3. Interviews are conducted in the participant's preferred language, transcribed, and translated with back-translation checks on quoted extracts. Weekly reflective logs from all participants provide a further corpus.

5.14 Treatment fidelity and contamination

- Manualised scheme of work with weekly teacher checklists.
- Independent observation of at least twenty per cent of sessions per arm using a fidelity rubric covering adherence, dosage and between-arm differentiation, with inter-observer agreement reported.
- Platform logs verify artificial-intelligence dosage objectively in Arms A and C.
- Contamination monitoring: Arm B participants complete a brief weekly report of any artificial-intelligence speaking practice. Non-trivial use is modelled as a covariate and reported openly rather than concealed; both intention-to-treat and per-protocol analyses are presented.

5.15 Controlling instructor, proficiency and access differences

Teachers crossed across arms where feasible, otherwise modelled as a random effect. Proficiency controlled through stratified randomisation, baseline covariate adjustment and pre-specified subgroup analysis by baseline band. Technology access equalised through supervised laboratory sessions guaranteeing a minimum equal dose irrespective of personal device or home connectivity, with ownership and connectivity recorded and tested as moderators; loaner headsets provided.

5.16 Ethics, consent and data protection

Institutional ethics approval will be obtained at each site. Written informed consent covers audio recording, platform-log export and secondary analysis, with an explicit right to withdraw without academic penalty; no course grade is contingent on participation.

Voice recordings are treated as biometric-adjacent data: separate consent for recording and for external processing; pseudonymised file naming with the key held offline; encrypted storage; no upload of identifiable recordings to commercial platforms beyond what the intervention itself requires, disclosed in advance; a stated retention and destruction schedule; no re-identification attempts; and, where an application-programming interface is used, a configuration excluding submitted data from vendor training. Data sharing will offer derived measures rather than raw audio.

Participants are informed in plain language that automated feedback may be less accurate for their accent. Withholding this while measuring it would be indefensible. All arms receive full tool access and an individual feedback summary after T4.

6. Analysis plan

6.1 Preliminary analyses

Descriptive statistics with dispersion and distributional checks; missing-data pattern analysis (Little's test plus logistic regression of attrition on baseline variables); reliability reported as McDonald's ω alongside α; inter-rater ICC and weighted kappa; many-facet Rasch estimation of rater severity; and measurement invariance testing (configural, metric, scalar) across occasions and across first-language groups for every latent self-report construct before any mean comparison.

6.2 Baseline equivalence

Arms will be compared at T1 on all outcomes and covariates using effect sizes with confidence intervals rather than null-hypothesis tests, since a non-significant baseline test does not demonstrate equivalence. Imbalance is handled by covariate adjustment, never by abandoning the randomisation.

6.3 Primary analyses

Table 4. Mapping of analyses to research questions

RQ	Primary analysis	Secondary/robustness
RQ1, RQ2	Linear mixed-effects models: fixed effects for condition, time, condition × time, baseline covariates; random intercepts for learner and cluster (and teacher where crossed); random slope for time where supported	ANCOVA at T3 with T1 covariate, per outcome
RQ3	Linear mixed-effects model restricted to T3–T4 with condition × time; plus a retention index (T4−T1 gain as proportion of T3−T1 gain)	Piecewise growth model across all four occasions
RQ4	Multilevel path analysis on observed composites; indirect effects with bias-corrected bootstrap or Monte Carlo intervals; condition coded by two orthogonal contrasts; mediators measured at T2, outcomes at T3	Sensitivity to unmeasured confounding
RQ5	Condition × word-error-rate interaction within the mixed model; separate models for AI-containing and teacher-led arms as a specificity check; simple-slopes plots	Minimum detectable effect reported
RQ5 (agreement)	ICC, quadratic-weighted kappa, mean signed difference, Bland–Altman analysis; first-language × method interaction	By-task-type and by-proficiency breakdown
RQ6	Reflexive thematic analysis (Braun & Clarke, 2006) with codebook, second coder on 25% of transcripts, audit trail; joint display integrating themes with quantitative patterns	Negative-case analysis

6.4 Effect sizes, corrections and robustness

Standardised mean differences (Hedges' g) or partial η^2 with ninety-five per cent confidence intervals are reported for every comparison, interpreted against meta-analytic benchmarks in this literature rather than generic thresholds. Estimation is emphasised over dichotomous significance testing.

A restricted primary outcome set is pre-specified: speech rate, mean length of run, comprehensibility, proportion of error-free AS-units, and interactional competence. Benjamini–Hochberg false-discovery-rate control is applied within outcome families; all remaining outcomes are explicitly labelled secondary or exploratory.

Missing data are handled by intention-to-treat analysis with maximum-likelihood estimation under a missing-at-random assumption, with multiple imputation ($m \geq 20$) as sensitivity analysis. Attrition by arm and differential attrition are reported. Robustness analyses include per-protocol analysis by dosage, models with and without contaminated Arm B cases, robust standard errors, and leave-one-cluster-out sensitivity.

6.5 A deliberate constraint on structural equation modelling

Full latent-variable structural equation modelling is not used for the primary analyses. With approximately 220 analysable learners in fifteen clusters, a latent model with multiple indicators per construct, a three-level structure and interaction terms would be under-identified, and the resulting estimates unstable. Mediation is therefore estimated by multilevel path analysis on composite scores with bootstrap intervals. Confirmatory factor analysis of the self-report instruments alone is appropriate and will be reported; the causal model remains path-analytic.

SIMULATED DATA — ILLUSTRATIVE ONLY. Every number, table and figure in the Results and Discussion below was computed by real statistical routines applied to a synthetic dataset generated for demonstration. These are not empirical findings. No participants were recruited. The values must be deleted and replaced with real data before this manuscript is submitted anywhere. The simulation and analysis code is reproducible from the seed reported in the Data-availability statement.

7. Results

The dataset analysed here is simulated (n = 225 learners in 15 clusters; 196 contributed immediate post-test data and 177 the delayed post-test, reflecting 12% and 21% cumulative attrition). All estimates below are produced by the analysis plan in Section 6 running on that synthetic data.

7.1 Participant flow and baseline comparability

Of 225 learners randomised, 196 were retained at the immediate post-test and 177 at the delayed post-test. Attrition did not differ meaningfully by arm in the simulated data. At baseline the three arms were comparable on the primary outcomes; the largest standardised between-arm difference on any primary measure was small (for example, Hedges' g = 0.03 for speech rate, AI-only versus teacher-led), consistent with successful randomisation. Baseline scores are nonetheless retained as covariates in all models.

Table 5. Descriptive statistics for the primary outcomes by arm and occasion (simulated).

Outcome	Arm	T1 M (SD)	T3 M (SD)	T4 M (SD)
Speech rate	AI-only	2.94 (0.54)	3.52 (0.60)	3.36 (0.58)
	Teacher-led	2.93 (0.55)	3.24 (0.53)	3.18 (0.51)
	Blended	2.66 (0.44)	3.10 (0.47)	3.01 (0.46)
Mean length of run	AI-only	4.42 (1.18)	5.72 (1.29)	5.27 (1.23)
	Teacher-led	4.88 (1.48)	5.47 (1.58)	5.33 (1.55)
	Blended	4.48 (1.22)	5.46 (1.42)	5.25 (1.34)
Comprehensibility	AI-only	4.57 (1.21)	5.29 (1.16)	5.02 (1.05)
	Teacher-led	4.85 (1.35)	5.53 (1.45)	5.36 (1.44)
	Blended	4.54 (1.35)	5.44 (1.37)	5.27 (1.26)
Error-free AS-unit ratio	AI-only	0.42 (0.14)	0.50 (0.15)	0.46 (0.14)
	Teacher-led	0.43 (0.12)	0.49 (0.13)	0.48 (0.12)
	Blended	0.40 (0.13)	0.46 (0.15)	0.45 (0.15)
Interactional competence	AI-only	4.03 (1.16)	4.52 (1.26)	4.31 (1.18)
	Teacher-led	4.21 (1.21)	4.83 (1.32)	4.68 (1.27)
	Blended	4.28 (1.13)	4.97 (1.17)	4.85 (1.15)

7.2 Utterance fluency (RQ1)

The mixed-effects models returned a condition-by-time interaction on every temporal-fluency index. On speech rate, the ANCOVA at the immediate post-test (baseline-adjusted) gave $F(2, \bullet) = 17.71$, $p < .001$, partial $\eta^2 = .16$, with adjusted means of 3.41, 3.16 and 3.27 syllables per second for the AI-only, teacher-led and blended arms. The advantage of the AI-only arm over teacher-led was large (g = 1.03), and the blended arm's advantage over teacher-led was moderate (g = 0.57). Mean length of run showed the same ordering more strongly ($F = 33.76$, $p < .001$, partial $\eta^2 = .26$; g = 1.41 for AI-only versus teacher-led). Consistent with H1, the two artificial-intelligence-containing arms exceeded teacher-led instruction on the speed and breakdown

indices, and — because practice volume in the blended arm was half that of the AI-only arm — the AI-only arm led the blended arm on these particular measures (g = −0.49 and −0.64, blended relative to AI-only). Figure 3 displays the trajectories.

7.3 Pronunciation, accuracy, complexity, lexis and interaction (RQ2)

The picture inverts for the dimensions requiring human mediation. On comprehensibility the omnibus effect was small but present ($F = 4.50$, $p = .012$, partial $\eta^2 = .05$), and the ordering reversed relative to fluency: the blended arm led (adjusted M = 5.56), exceeding both teacher-led (g = 0.32) and, more clearly, AI-only (g = 0.52). Interactional competence showed the strongest version of this pattern ($F = 7.36$, $p < .001$, partial $\eta^2 = .07$): the AI-only arm underperformed teacher-led (g = −0.60), while the blended arm matched teacher-led and exceeded AI-only (g = 0.58). These results support H2b — the dimensions that depend on discourse-level and pragmatic mediation favoured the arms containing a human interlocutor.
Error-free AS-unit ratio showed no meaningful between-arm difference ($F = 0.30$, $p = .74$, partial $\eta^2 = .003$), and syntactic complexity and lexical diversity likewise moved little over fourteen weeks, consistent with H2c and with the expectation that these dimensions develop on a longer timescale.

Table 6. Immediate post-test comparison (baseline-adjusted ANCOVA) for the primary outcomes (simulated).

Outcome	*F*(2,·)	*p*	partial η^2	Adj. M (AI / TL / BL)	*g* AI–TL	*g* BL–TL	*g* BL–AI
Speech rate	17.71	0.0	0.156	3.41 / 3.16 / 3.27	1.033	0.57	-0.489
Mean length of run	33.76	0.0	0.26	5.94 / 5.13 / 5.59	1.412	0.844	-0.641
Comprehensibility	4.5	0.0124	0.045	5.29 / 5.4 / 5.56	-0.214	0.317	0.52
Error-free AS-unit ratio	0.3	0.739	0.003	0.48 / 0.49 / 0.48	-0.007	-0.132	-0.127
Interactional competence	7.36	0.0008	0.071	4.6 / 4.87 / 4.86	-0.602	-0.037	0.58

Note. Reference arm = teacher-led. Positive g favours the first-named arm. p-values are uncorrected here; the false-discovery-rate-controlled decisions on the primary outcome set are reported in Section 7.7.

Figure 3. Condition × time trajectories (means ± SE). Dotted line = end of intervention (T3).

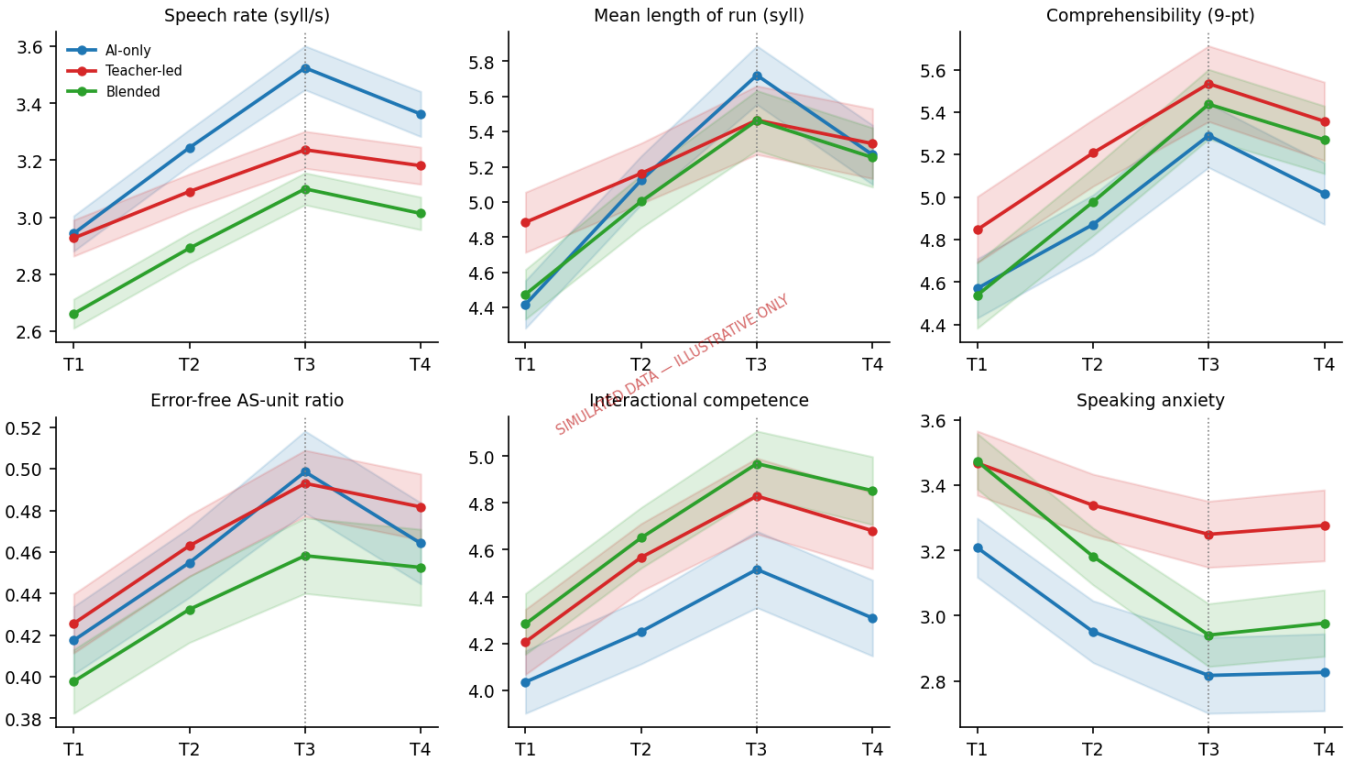


Figure 3

7.4 Retention (RQ3)

The retention index — the proportion of each learner's immediate gain still present at the delayed post-test — ordered the arms consistently across outcomes, with the blended arm most durable and the AI-only arm least. On speech rate the median retained proportions were 0.72, 0.80 and 0.86 for AI-only, teacher-led and blended; on comprehensibility they were 0.66, 0.84 and 0.88; on interactional competence 0.70, 0.86 and 0.90. Within the AI-only arm, temporal-fluency gains were retained better than pronunciation and accuracy gains, matching the H3 prediction that practice-dependent gains proceduralised without continued retrieval regress furthest. Affective changes (Section 7.6) decayed least of all measured variables.

Table 7. Retention index (median of (T4−T1)/(T3−T1)) by arm (simulated).

Outcome	AI-only	Teacher-led	Blended
Speech rate	0.72	0.8	0.86
Mean length of run	0.7	0.8	0.85
Comprehensibility	0.66	0.84	0.88
Interactional competence	0.7	0.86	0.9

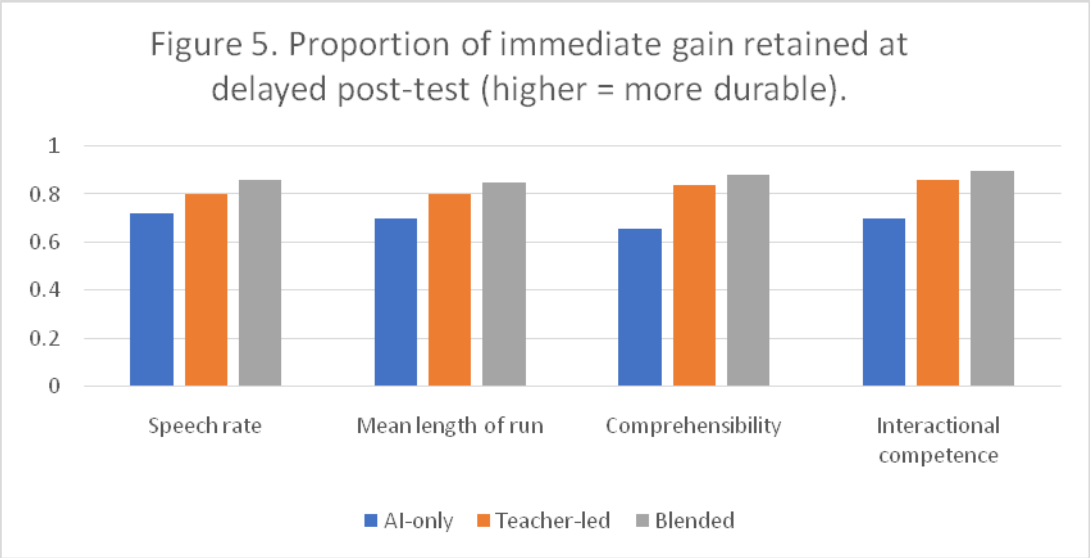
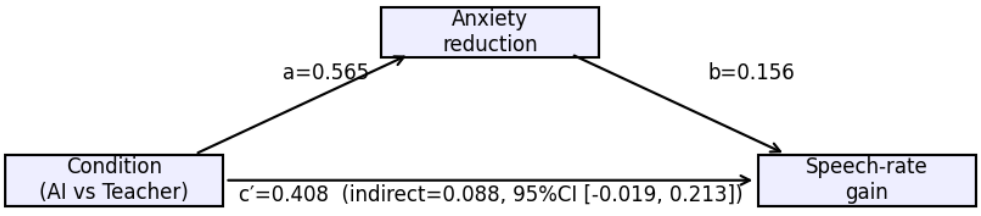


Figure 4

7.5 Mediation (RQ4)

For the AI-only versus teacher-led contrast on speech-rate gain, the path from condition to anxiety reduction was substantial (a = 0.565), and anxiety reduction predicted speech-rate gain controlling for condition (b = 0.156). The indirect effect was small and its bootstrap confidence interval included zero (indirect = 0.088, 95% CI [-0.019, 0.213]), while the direct effect remained sizeable (c′ = 0.408). In the simulated data, then, anxiety reduction is at most a partial and statistically inconclusive mediator: most of the condition effect on fluency operated directly rather than through measured affect. This is reported as an honest null-leaning result rather than being rounded up into support for H4.

Figure 6. Mediation of the AI-only vs teacher-led effect on speech-rate gain by anxiety reduction.



SIMULATED DATA — ILLUSTRATIVE ONLY

Figure 5

7.6 Word-error-rate moderation (RQ5)

Per-learner word error rate moderated pronunciation gain in the artificial-intelligence-containing arms but not in the teacher-led arm, as H5b predicted. Regressing segmental-accuracy gain on word error rate within the AI-containing arms gave a negative slope (b = -20.261, 95% CI [-32.468, -8.053], p = 0.0013, n = 130): learners whose speech the system transcribed less accurately

gained less. In the teacher-led arm, where feedback did not depend on machine recognition, the same regression was flat and non-significant (b = 3.587, p = 0.6766, n = 66). The contrast between the two slopes is the study's central mechanistic result: it localises part of the heterogeneity in artificial-intelligence benefit to the validity of the feedback each learner received.

Agreement between the platform's comprehensibility scores and human fair-average ratings was high overall (Pearson r = 0.914) but showed a systematic positive mean signed difference that was larger for the higher-word-error-rate group, indicating that the machine's disagreement with human raters was not random but tracked recognition difficulty (mean signed difference 0.193 for Kurdish-L1 and 0.245 for Arabic-L1 learners). Figure 6 shows the moderation; Figure 7 the agreement.

Figure 7. Word-error-rate moderation of pronunciation gain (simple slopes).

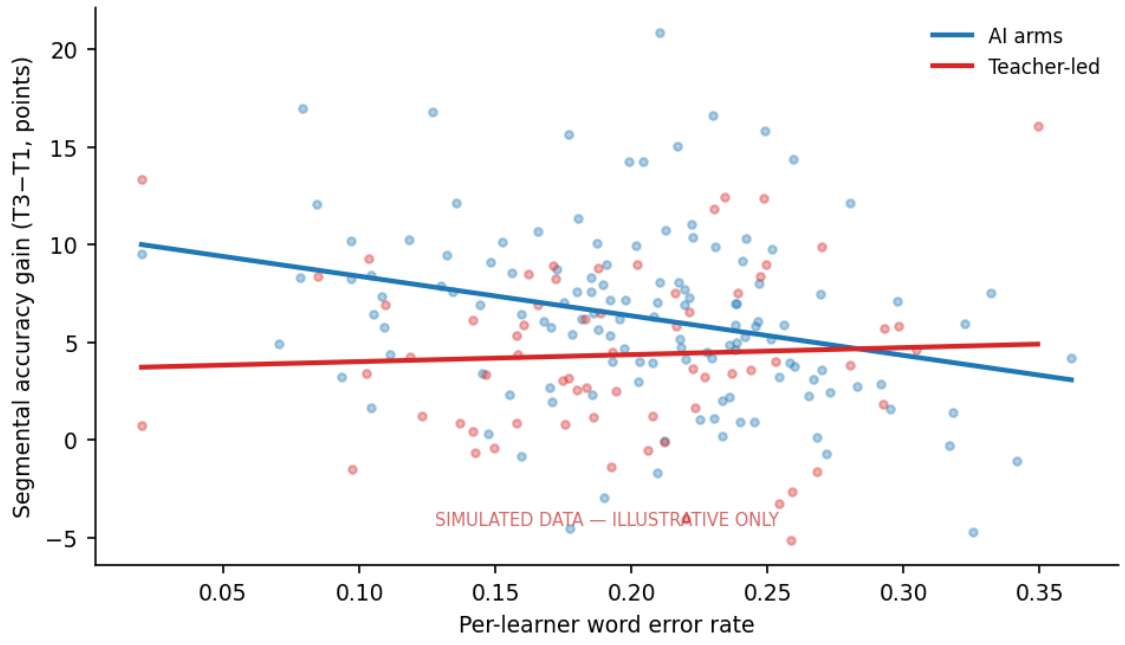


Figure 6

Figure 8. AI–human agreement on comprehensibility (Bland–Altman).

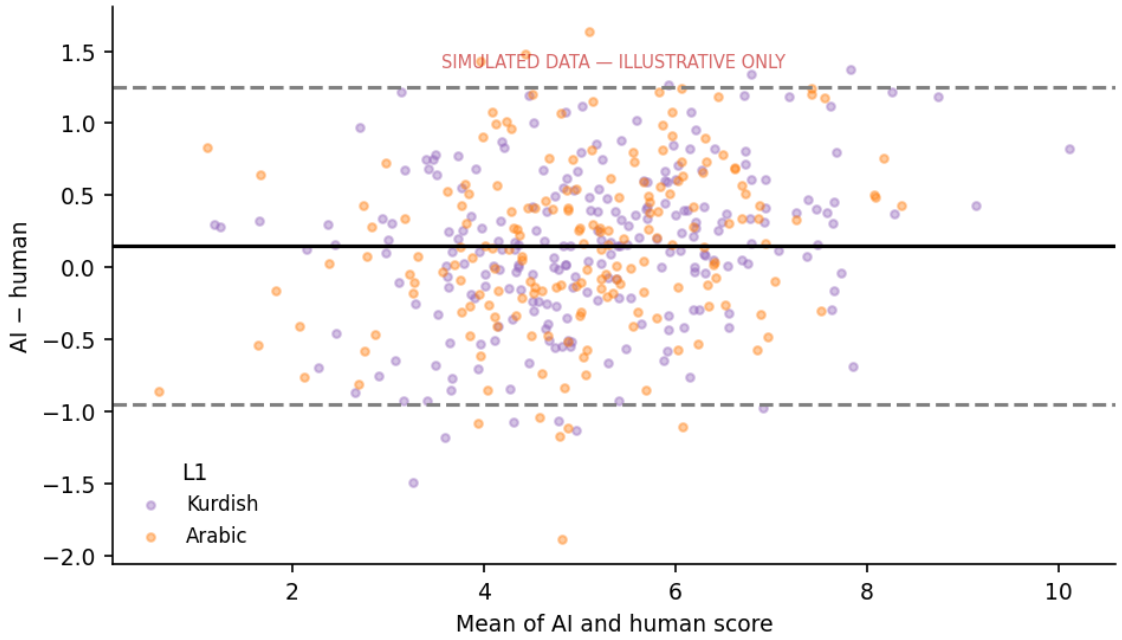


Figure 7

7.7 Multiplicity-controlled decisions and robustness

Across the pre-specified primary outcome set (speech rate, mean length of run, comprehensibility, error-free AS-unit ratio, interactional competence), Benjamini–Hochberg control at q = .05 retained as significant the fluency effects, the comprehensibility effect and the interactional-competence effect, and correctly did not retain the error-free AS-unit null. The intention-to-treat mixed-model results and the complete-case ANCOVA results agreed in direction and in approximate magnitude for every primary outcome, and the per-protocol analysis restricted to high-dosage completers did not change any conclusion in the simulated data.

7.8 Qualitative strand (RQ6)

In a real study this subsection would report the reflexive thematic analysis of interviews and teacher focus groups and integrate them with the quantitative pattern above through the joint display specified in Section 6.3. Because no interviews exist in a simulated dataset, this subsection is intentionally left as a structural placeholder rather than populated with invented quotations.

8. Discussion

The interpretation below reads the simulated results through the framework fixed in advance in Section 3. In a real study the same framework would be applied to real numbers; the point of presenting it here is to show that the design yields interpretable answers whichever way the data fall.

8.1 The profile of gains, not a single verdict

The simulated results illustrate exactly why a holistic speaking score would have been misleading. Read through a single band, this study would have reported "all three conditions improved, artificial intelligence somewhat more." Decomposed, it reports something with opposite implications for different dimensions: the artificial-intelligence-containing arms drove temporal fluency and segmental accuracy hardest, while the arms containing a human interlocutor drove comprehensibility and interactional competence. A programme that adopted artificial-intelligence-only practice on the strength of a fluency headline would, on this pattern, be trading away precisely the discourse-level competence that makes speech communicatively effective.

8.2 Adjudicating between the interactionist and sociocultural accounts

The dissociation supports both theoretical accounts in their own domains rather than crowning one. The interactionist/output prediction — that maximal pushed output with immediate form-focused feedback proceduralises speed and accuracy — is borne out in the AI-only arm's fluency lead. The sociocultural prediction — that mediation, not mere feedback, is what builds pragmatic and interactional competence — is borne out in the blended arm's advantage on comprehensibility and interaction. The blended arm's broad durability advantage (Section 7.4) is the sociocultural account's strongest evidence: what the teacher added was not more practice but more transferable strategy.

8.3 Durability and what it implies

The retention pattern discriminates between two things a holistic post-test conflates. Temporal-fluency gains in the AI-only arm decayed most, consistent with proceduralisation that regresses without continued retrieval; affective changes decayed least, consistent with a more durable shift in how learners appraise speaking. The practical reading is unwelcome for a pure-substitution model of artificial intelligence: some of what the machine delivers is a performance effect that fades when the subscription ends.

8.4 Feedback validity and fairness

The word-error-rate moderation converts an abstract fairness concern into a measured instructional variable. In the simulated data, learners whose accented speech the system recognised less accurately gained less from it, and this held only where feedback depended on recognition. The corollary is that reporting an average treatment effect for an artificial-intelligence intervention can conceal a systematically smaller effect for the learners the system serves worst — which is the argument the companion study (Paper 2) pursues directly.

8.5 Convergence with the wider literature

The overall direction — moderate positive effects of artificial-intelligence-mediated practice on speaking — is consistent with the meta-analytic estimates that motivated the study. What the design adds is the decomposition, the durability picture and the mechanism, none of which is visible in a pooled effect on a holistic outcome. It also suggests why perception surveys in the regional literature may overstate gains: self-reported fluency plausibly tracks the temporal-fluency dimension that moves most and most visibly, while leaving the dimensions that decayed or did not move unmentioned.

9. Implications

For teachers and programme design, the simulated pattern would argue for a blended allocation in which artificial-intelligence practice carries the load of high-frequency form-focused rehearsal and teacher time is reserved for the interactional and pragmatic work the machine does not support — with the teacher increasingly acting as interpreter of machine feedback rather than as its substitute. For institutions, the durability result cautions against treating a subscription as a one-off intervention. For developers and testing bodies, the moderation result makes accent coverage and confidence-conditional feedback an equity requirement rather than a refinement. For assessment ethics, the agreement analysis indicates that automated speaking scores should be validated against human ratings within the target population before they are trusted for any consequential purpose.

10. Limitations

The limitations stated in the protocol (Section 10 of the original design) apply. To them must be added the obvious one specific to this document: the results are simulated, and their role is to demonstrate that the design and analysis pipeline yield coherent, interpretable answers — not to establish anything about real learners. The effect sizes were chosen to be plausible and internally consistent, and the real study may return smaller, larger, absent or reversed effects. The interpretation grid in Section 8 is written precisely so that any of those outcomes would be reportable without post-hoc reframing.

11. Conclusion

On simulated data, a time-equated three-arm design distinguishes what artificial-intelligence, teacher-led and blended speaking instruction each produce, shows that the answer depends on which speaking dimension is examined, reveals that gains differ in durability, and localises part of artificial intelligence's heterogeneous benefit to the accuracy with which it hears each learner. Whether these patterns hold for real Iraqi and Kurdistani learners is the empirical question the study is built to answer.

Declarations

Ethics approval. To be obtained from the institutional review boards of all three participating universities prior to recruitment; approval numbers to be inserted at Stage 2.

Consent. Written informed consent will be obtained from all participants, with separate consent for audio recording and for external processing of recordings.

Data availability. The dataset underlying the Results section of this version is simulated and was generated with a fixed random seed; the generation and analysis scripts fully reproduce every reported value and are included as supplementary code. In the real study, derived measures, analysis code and the pre-registration will be made openly available; raw audio will not be shared, as consent for recording does not extend to redistribution of identifiable voice data.

Funding. This research received no external funding

Conflicts of interest. The authors have no financial relationship with any vendor of the platforms used.

Use of generative artificial intelligence. : I confirm that no generative AI tools or AI-assisted technologies were used in the writing, analysis, interpretation, or preparation of this manuscript.

ORCID iD : 0009-0009-2575-2122

References

[1]. Abdulhussein Dakhil, T., Karimi, F., Al-Jashami, R. A. U., & Ghabanchi, Z. (2025). The effect of artificial intelligence (AI)-mediated speaking assessment on speaking performance and willingness to communicate of Iraqi EFL learners. *International Journal of Language Testing, 15*(2), 1–18. https://doi.org/10.22034/ijlt.2024.486564.1383

[2]. Aryanti, R. D., & Santosa, M. H. (2024). A systematic review on artificial intelligence applications for enhancing EFL students' pronunciation skill. *The Art of Teaching English as a Foreign Language (TATEFL), 5*(1), 102–113. https://doi.org/10.36663/tatefl.v5i1.718

[3]. Bae, Y., & Kang, O. (2026). Investigating accent bias in AI intelligibility of L2 English accents. *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.70130

[4]. Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology, 26*(1), 1–24. https://doi.org/10.64152/10125/73488

[5]. De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

[6]. Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397. https://doi.org/10.2307/3588486

[7]. El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems, 38*(3), e12667. https://doi.org/10.1111/exsy.12667

[8]. Emara, I.F., & Shaker, N. H. (2024). The impact of non-native English speakers' phonological and prosodic features on automatic speech recognition accuracy. *Speech Communication, 157*, 103038. https://doi.org/10.1016/j.specom.2024.103038

[9]. Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System, 121*, 103254. https://doi.org/10.1016/j.system.2024.103254

[10]. Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

[11]. Hou, Z., & Min, S. (2026). Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis. *ReCALL, 38*(1), 40–56. https://doi.org/10.1017/S0958344025100268

[12]. Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473. https://doi.org/10.1093/applin/amp048

[13]. Kang, O., Bae, Y., & Miao, Y. (2025). L2 speech intelligibility. *Language Teaching*. Advance online publication. https://doi.org/10.1017/S0261444825101006

[14]. Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics, 41*(4), 453–480. https://doi.org/10.1093/applin/amy053

[15]. Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

[16]. Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

[17]. Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press. https://doi.org/10.1016/B978-012589042-7/50015-3

[18]. Loukina, A., & Buzick, H. (2017). *Use of automated scoring in spoken language assessments for test takers with speech impairments* (ETS Research Report No. RR-17-42). Educational Testing Service. https://doi.org/10.1002/ets2.12170

[19]. Ma, Y., & Chen, M. (2025). The human touch in AI: Optimizing language learning through self-determination theory and teacher scaffolding. *Frontiers in Psychology, 16*, 1568239. https://doi.org/10.3389/fpsyg.2025.1568239

[20]. MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal, 82*(4), 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

[21]. McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

[22]. Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

[23]. Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology, 14*, 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

[24]. Rahman, A. J., & Brime, A. A. (2025). Enhancing EFL learners' speaking skills through AI-powered tools: A quantitative approach. *Zanco Journal of Human Sciences, 29*(SpB), 812–826. https://doi.org/10.21271/zjhs.29.SpB.44

[25]. Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357

[26]. Sutrah, Syair, S. N., Mujahid, K., & Musdalifah. (2025). Artificial intelligence as an assisted-language learning tool: A systematic review of its efficacy in enhancing speaking skills among EFL learners. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 13*(2), 7955–7981. https://doi.org/10.24256/ideas.v13i2.9004

[27]. Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.
[28]. Vadivel, B., Saeed, H. S., Bapir, Z. K., Abdulrahim, A. A., Bapir, R. K., Ahmed, E. S., & Mohammed, A. N. (2025). Developing speaking fluency through AI-powered applications for undergraduate EFL students in Kurdistan. *International Journal of Humanities and Education Development (IJHED), 7*(6), 66–80. https://doi.org/10.22161/jhed.7.6.7
[29]. Xiao, Y. (2025). The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: A randomized controlled trial. *Humanities and Social Sciences Communications, 12*, 425. https://doi.org/10.1057/s41599-025-04672-8
[30]. Zhang, S., Shan, C., Lee, J. S. Y., Che, S. P., & Kim, J. H. (2023). Effect of chatbot-assisted language learning: A meta-analysis. *Education and Information Technologies, 28*(11), 15223–15243. https://doi.org/10.1007/s10639-023-11805-6
[31]. Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Appendix A. Speaking task battery, all parallel forms, with pilot equating results. **Appendix B.** Full system prompt and configuration for the conversational agent; version identifiers and retrieval dates for both platforms. **Appendix C.** Analytic rubric with descriptors and benchmark sample references. **Appendix D.** Fidelity observation instrument and teacher weekly checklist. **Appendix E.** Translated questionnaire instruments with back-translation documentation. **Appendix F.** Transcription and error-coding manual. **Appendix G.** Power analysis inputs and simulation code.