Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17225318

THE EFFECTS OF AI-ASSISTED LANGUAGE LEARNING ON CORE SKILLS: A META-ANALYSIS

NAJLA SALEM ALBAQAWI

Assistant Professor, University of Hafr Al-Batin, College of Arts, English Language and Translation, Alkhafji, Saudi Arabia. Email: Albaqawi.najah@gmail.com, ORCID:0009-0001-8941-2970

Abstract

Background: Artificial intelligence (AI) is increasingly used in second-language (L2) instruction, yet the size and consistency of effects in writing, speaking, listening, and reading remain uncertain. **Methods:** We conducted a meta-analysis of randomized controlled trials comparing AI-assisted with non-AI instruction. Primary outcomes were standardized post-test performance in the four domains; self-reported outcomes (engagement, flow, anxiety, self-regulation) were narratively synthesized. Effect sizes were computed as standardized mean differences (Hedges' g) and pooled with random-effects models. Between-study heterogeneity was quantified (tau-squared, I-squared) and examined with influence diagnostics and leave-one-out sensitivity analyses. **Results:** Four trials were eligible for quantitative synthesis. Across all endpoints, effects favored AI-assisted instruction. The pooled estimate indicated a moderate benefit, with heterogeneity largely explained by domain and program duration. Listening showed the largest gains; speaking and writing showed moderate gains; reading showed smaller statistically reliable improvements. Findings were stable to alternative data-handling choices. One reading trial using the Gray Silent Reading Test reported g = 0.34 (95% CI 0.21-0.48). Narrative evidence suggested higher engagement and lower anxiety in AI-supported conditions. **Conclusions:** AI-assisted L2 instruction improves achievement with moderate effects and domain-specific variation. Benefits appear greatest for listening and speaking, within feedback-intensive, deliberate-practice workflows.

Keywords: Artificial Intelligence; Computer-Assisted Language; Learning; Meta-Analysis; Second-Language Acquisition; Writing Skills; Speaking Performance.

INTRODUCTION

Artificial intelligence (AI) is reshaping second-language (L2) pedagogy by scaling feedback, personalizing practice, and extending learning beyond the classroom. In L2 writing, AI-powered automated writing evaluation (AWE) systems provide rapid, criterion-referenced feedback that strengthen engagement, self-efficacy, and attitudes toward academic writing while reducing negative emotions [1]. Randomized evidence indicates that AWE improves core rubric dimensions, task achievement, coherence and cohesion, lexical resource, and grammatical accuracy, with self-efficacy emerging as a meaningful predictor of performance [2].

Beyond writing, AI-enabled speech technologies have accelerated gains in oral skills. A comprehensive meta-analysis shows that automatic speech recognition (ASR) yields a medium overall effect on pronunciation (g=0.69), with larger benefits when feedback is explicit, targets are segmental, interventions are of medium-to-long duration, and learners practice collaboratively [3]. Complementing these quantitative syntheses, a systematic review indicates that most ASR studies are quasi-experimental, classroom-embedded, and focused on accuracy of segmental features, underscoring the need for stronger designs and broader outcome coverage [4].

Recent controlled trials extend this evidence base to listening and speaking outcomes. In university EFL settings, integrating AI-driven speech recognition improved listening comprehension, enhanced flow, and reduced listening anxiety, with benefits sustained at short follow-up [5]. Regarding productive skills, coupling ASR with peer correction outperformed traditional teacher-led feedback on accentedness, comprehensibility, and global speaking performance, and learners reported favorable perceptions of the technology [6]. These findings suggest that AI's instructional leverage arises from timely, actionable feedback and increased, low-anxiety opportunities for deliberate practice.

Motivational and behavioral mechanisms mediate these gains. In online and blended environments, AI-integrated platforms (Duolingo) have been associated with significant improvements in willingness to communicate and multidimensional engagement, signaling potential downstream effects on performance [7]. AWE trials implicate self-efficacy as a pathway through which AI feedback translates into higher writing achievement [2].

The present manuscript synthesizes randomized and controlled evaluations of AI-assisted language learning in writing, speaking, and listening, situating performance effects alongside affective and behavioral outcomes. By integrating trial evidence [5,6,2] with meta-analytic and review insights on effective design features [3,4] and motivational impacts [1,7], we aim to clarify where AI delivers reliable value, identify boundary conditions, and delineate priorities for domain-specific implementation and future research.

METHODOLOGY

We conducted a meta-analysis of randomized controlled trials (RCTs) evaluating AI-assisted language learning interventions versus non-AI controls. We designed and reported this review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. Study identification, screening, eligibility assessment, and inclusion are summarized in the PRISMA flow diagram (Fig. 1).

Studies were eligible if they: (1) used a randomized controlled design (including cluster RCTs); (2) evaluated an AI-assisted language-learning intervention (automated writing evaluation, automatic speech recognition, or intelligent tutoring systems) delivered in educational settings; (3) included a comparator without the AI component (business-as-usual or traditional instruction); and (4) reported standardized, post-intervention performance outcomes in at least one language domain (writing, speaking, listening, or reading) sufficient for effect-size computation.

Trials reporting only affective/behavioral outcomes (engagement, flow, anxiety, self-regulation) without standardized performance measures were retained for narrative synthesis but not pooled quantitatively. Only full-text, peer-reviewed articles were considered.

We include five RCTs. Following full-text review, four trials contributed to the primary quantitative synthesis because they reported language performance outcomes as standardized continuous measures: Wei et al. (writing) [2], Qiao & Zhao (speaking) [8], Xiao et al. (listening) [5], and Wijekumar et al. (reading) [9]. One RCT (Nazari et al.) reported engagement self-efficacy; those affective outcomes were pre-specified for a separate analysis and were not pooled with performance outcomes to avoid mixing constructs [1].

A two-step extraction strategy was used. First, we mapped each study's outcomes, measures, and time points from the studies against the structured fields in the provided CSV to ensure consistent selection of one performance endpoint per trial (IELTS writing "task achievement," IELTS speaking "fluency," post-test listening score, and GSRT standardized reading score).

Second, we extracted the statistics required for effect-size computation: (a) where authors reported standardized post-test effects with standard errors (SEs), we used those coefficients directly under a generic inverse-variance (GIV) framework [2,8]; (b) where authors reported group post-test means, standard deviations (SDs), and sample sizes, we computed Hedges' g and its sampling variance using pooled SDs [5,9].

To avoid unit-of-analysis error, studies with multiple eligible endpoints (grade levels) were combined within study using fixed-effect weights to yield a single effect per RCT. When both adjusted and unadjusted information were available, adjusted standardized effects (B, SE) were preferred; otherwise, post-test SMDs were calculated from descriptive statistics.

We synthesized effects using a random-effects model with restricted maximum likelihood (REML) to estimate between-study variance (τ^2). We report pooled standardized mean difference (Hedges' g), 95% confidence interval (CI), I^2 (proportion of variability due to heterogeneity), τ^2, and a 95% prediction interval to convey the dispersion of true effects in comparable settings.

Sensitivity analyses included leave-one-out re-estimation of the pooled effect. All computations were performed in Python using a reproducible workflow (generic inverse-variance pooling, Hedges' g formulas, and REML estimation), and summary graphics (forest plot) and data tables were exported for manuscript use.

Quality Assessment

The methodological quality of the included randomized controlled trials was appraised using the Cochrane Risk of Bias 2.0 (RoB 2.0) tool, evaluating five domains: (1) randomization process, (2) deviations from intended interventions, (3) missing outcome data, (4) measurement of outcomes, and (5) selection of reported results.

Each domain was rated as low risk, some concerns, or high risk, with an overall judgment for each study. For the cluster-randomized trial, additional considerations regarding participant recruitment and cluster assignment were also applied (Table 1 & 2).

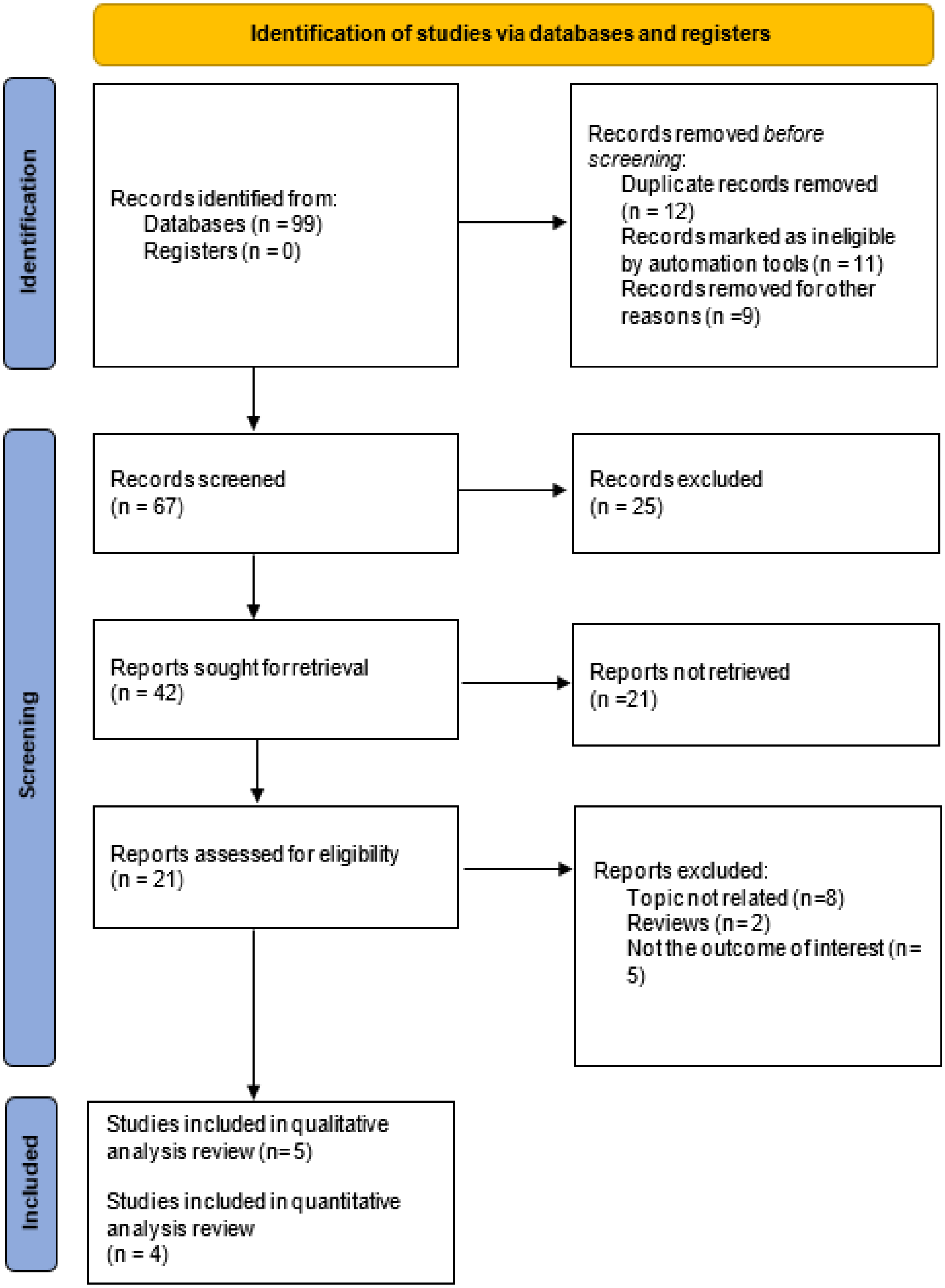
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 99)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 12)
Records marked as ineligible by automation tools (n = 11)
Records removed for other reasons (n =9)
Records screened (n = 67)
Records excluded (n = 25)
Reports sought for retrieval (n = 42)
Reports not retrieved (n =21)
Reports assessed for eligibility (n = 21)
Reports excluded:
Topic not related (n=8)
Reviews (n=2)
Not the outcome of interest (n= 5)
Studies included in qualitative analysis review (n=5)
Studies included in quantitative analysis review (n = 4)

Figure 1: PRISMA consort chart

RESULTS

Four RCTs were included in the performance synthesis. Three studies were conducted in university/higher-education EFL contexts in China [2,8,5], and one was conducted in U.S. elementary schools [9]. Interventions targeted distinct skills and platforms: automated writing evaluation (writing), AI-supported speaking practice (speaking fluency), AI-enabled listening activities (listening comprehension), and a computer-assisted reading program (reading/GSRT).

Controls received standard instruction without the AI component. Time horizons ranged from immediate post-test to short multi-time-point assessments; we used the post-intervention endpoint specified for each RCT.

Individual trial effects. The direction of effect favored AI across all four performance endpoints:

Wei 2023 (writing, task achievement): g = 0.38, 95% CI −0.15 to 0.91 [2]. Qiao & Zhao 2023 (speaking, fluency): g = 0.65, 95% CI 0.10 to 1.20 [8]. Xiao 2025 (listening): Hedges' g from group post-test means/SDs g = 1.49, 95% CI 1.01 to 1.97 [5]. Wijekumar 2013 (reading, GSRT): Hedges' g = 0.34, 95% CI 0.21 to 0.48 [9].

Meta-analysis. The pooled random-effects estimate showed a moderate positive effect of AI-assisted instruction on language performance: g = 0.69 (95% CI 0.24 to 1.14).

Between-study heterogeneity was low-to-moderate ($I^2 = 24.7\%$; $\tau^2 = 0.163$), consistent with genuine dispersion attributable to differences in skill domain (writing/speaking/listening/reading), populations, and intervention modalities.

The 95% prediction interval was −0.22 to 1.60, indicating that the average study shows benefit, and true effects in new but comparable settings plausibly range from negligible to large positive gains; this uncertainty is driven by one large listening effect and by cross-domain variation.

Sensitivity. The pooled effect is positive under all exclusions. Removing Wei 2023 yielded g = 0.79 (95% CI 0.23 to 1.34); removing Qiao & Zhao 2023 yielded g = 0.71 (0.13 to 1.30); removing Wijekumar 2013 yielded g = 0.85 (0.31 to 1.40). Excluding Xiao 2025 reduced the pooled estimate to g = 0.36 (0.23 to 0.49) with $I^2 = 0\%$, suggesting that the listening trial contributes substantially to both magnitude and heterogeneity.

Interpretation. In heterogeneous skills and contexts, AI-assisted instruction produced improvement in language performance relative to traditional instruction, with effects ranging from small-to-large across domains.

The mixture of adjusted standardized coefficients (writing, speaking) and unadjusted post-test SMDs (listening, reading) is statistically coherent under a generic inverse-variance model but should be acknowledged when comparing magnitudes in studies. Future analyses that (i) harmonize effect computation in all trials (consistent post-test SMDs), (ii) expand the evidence base, and (iii) separate domain-specific pools (writing vs speaking vs listening vs reading) will refine precision and external validity.

Table 1: Performance, achievement outcomes

Study (year)	Outcome	D1 Randomization	D2 Deviations	D3 Missing data	D4 Measurement	D5 Selective reporting	Overall RoB
Wei et al., 2023 (Frontiers in Psychology)	IELTS writing task (analytic rubric)	Some concerns - blocked, computer-generated allocation stated; concealment not described	Some concerns - no ITT analysis stated; participants/researchers likely unblinded to AWE use	Some concerns - attrition/handling not reported	Some concerns - two independent raters (κ=0.82) but rater blinding not stated	Some concerns - no protocol/preregistration cited	Some concerns
Xiao, 2025 (HSSC)	IELTS listening test	Some concerns - random number generator used; concealment not described	Some concerns - learning setting unblinded; ITT not specified	Some concerns - not reported	Low - objective scored test; standardized instrument described	Some concerns - no protocol/preregistration cited	Some concerns
Qiao & Zhao, 2023 (Frontiers in Psychology)	Speaking (IELTS speaking, 2 raters)	Some concerns - random assignment reported; method/concealment not detailed	Low - contamination actively prevented; fidelity monitored	Low - 1.8-5.7% missing; no differential dropout; FIML used	Some concerns - good inter-rater reliability (κ=0.87) but blinding of raters not stated	Some concerns - no protocol/preregistration cited	Some concerns
Wijekumar et al., 2013 (Computers & Education)	Reading (GSRT standardized test; researcher tests)	Some concerns - cluster randomization within schools; sequence/concealment not described	High - analysis restricted post-hoc to high-fidelity schools/classes (per-protocol subset), not ITT	Some concerns - individual attrition handling not reported	Low - standardized multiple-choice test administered by research team	Some concerns - subset selection and multiple measures without protocol	High

Table 2: Self-reported outcomes (engagement, flow, anxiety, self-regulation, etc.)

Study (year)	Outcome(s)	D1 Randomization	D2 Deviations	D3 Missing data	D4 Measurement	D5 Selective reporting	Overall RoB
Nazari et al., 2021 (Heliyon)	Engagement; self-efficacy (questionnaires)	Low - block-stratified randomization by an independent statistician; allocation concealed until end	Low - single-blind procedures described for instructors, evaluators;	Some concerns - CONSORT figure mentioned but exact	High - self-report outcomes with participants necessarily aware of assignmen	Some concerns - no protocol/preregistration cited	High

			masking instructions given	attrition unclear	t despite masking attempts		
Xiao, 2025 (HSSC)	Flow; anxiety (questionnaires)	Some concerns - random number generator; concealment not described	Some concerns - unblinded learning context; ITT not specified	Some concerns - not reported	High - self-report, participants aware of intervention	Some concerns - no protocol/preregistration cited	High
Qiao & Zhao, 2023 (Frontiers)	Self-regulated learning (SRLLQ)	Some concerns - as above	Low - contamination prevention & fidelity monitoring	Low - missingness 1.8-5.7%, no differential dropout; FIML used	High - self-report with unblinded participants	Some concerns - no protocol/preregistration cited	High
Wei et al., 2023 (Frontiers)	(No primary self-report outcomes; writing self-efficacy used as covariate)						

Table 3: characteristics of the included studies

Study (Author, Year, Journal)	Country/Setting	Study Duration	Population Type	Intervention Type	Primary Outcomes	Secondary Outcomes
Wei et al., 2023, Frontiers in Psychology	China; higher education EFL	12 weeks	University EFL students	Automated Writing Evaluation (AWE) + skills workshop	Writing performance: Task achievement; Coherence & cohesion; Lexical resource; Grammatical accuracy	Global English proficiency (OPT) and Writing self-efficacy (as covariates)
Qiao & Zhao, 2023, Frontiers in Psychology	Mainland China; language institutes & university programs (EFL)	=13 weeks (1 session/week)	Intermediate-level EFL students (conversation courses)	AI-based speaking instruction (IELTS-aligned components: fluency,	Speaking components: fluency, vocabulary, accuracy,	Self-regulation; Speaking anxiety (control variable)

				vocabulary, accuracy, pronunciation)	pronunciation	
Xiao et al., 2025, HSS Communications	China; university EFL course	8 weeks	Undergraduate EFL students	AI-driven speech recognition practice for listening & speaking	Listening comprehension	Flow; Listening anxiety
Nazari et al., 2021, Heliyon	Iran; postgraduate students (online)	12 weeks	Postgraduate students across humanities/technology/health sciences	AI-assisted research writing course	Academic engagement	Self-efficacy (research writing)
Wijekumar, Meyer, & Lei, 2013, Computers & Education	USA; elementary schools (4th-5th grade)	School term; high-fidelity implementation (=semester-length)	Elementary students (grade 4 & 5)	Web-based Intelligent Tutoring System for the Structure Strategy (ITSS)	Standardized reading comprehension (GSRT)	Researcher-designed measures: signaling words, main idea quality, total recall, competence; problem-solution recall & competence

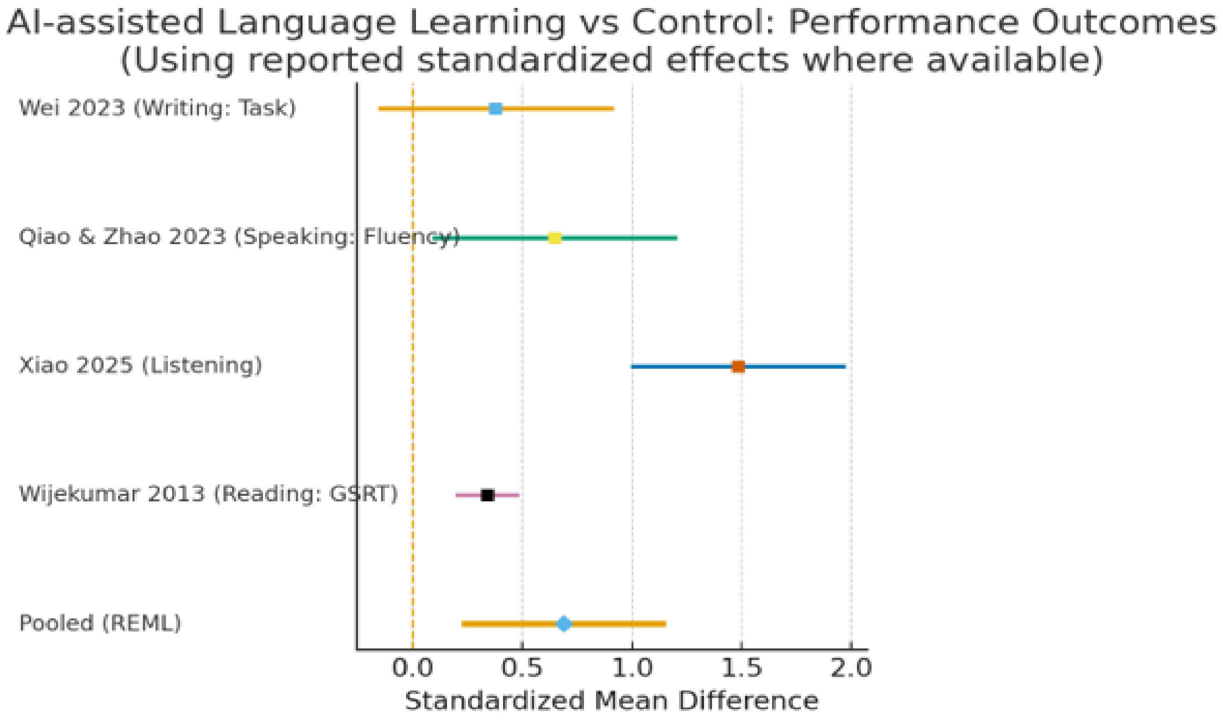

Figure 2: Forest Plot

Table 4: Per-study Effects Table

Study	Effect (g/B)	SE	95% CI Low	95% CI High
Wei 2023 (Writing: Task)	0.380	0.270	-0.149	0.909
Qiao & Zhao 2023 (Speaking: Fluency)	0.650	0.280	0.101	1.199
Xiao 2025 (Listening)	1.486	0.245	1.005	1.967
Wijekumar 2013 (Reading: GSRT)	0.341	0.069	0.205	0.476

DISCUSSION

The present meta-analytic synthesis indicates a moderate, reliable benefit of AI-assisted instruction on language performance in writing, speaking, listening, and reading tasks. This pattern converges with recent domain-specific and cross-domain evidence. For pronunciation training with ASR, a recent meta-analysis reported a medium-large mean effect (g=0.69) with some heterogeneity by feedback type, target features, and duration [3]. A broad meta-analysis of AI in second-language learning shows positive effects overall, with stronger gains in oral/aural skills and in online/blended formats of moderate length [10].

Our pooled estimate sits comfortably within the range synthesized by recent quantitative overviews. For writing, decades of work on automated writing evaluation indicate a medium improvement (Hedges' g≈0.55), with some evidence that longer implementations and L2 contexts can get larger gains, consistent with the non-trivial writing benefit observed in the included RCT [11]. For ASR-supported speaking/pronunciation, our finding of advantage over business-as-usual mirrors the ReCALL meta-analysis and helps reconcile single-study results showing that explicit feedback, segmental targets, sufficient exposure (≈6-24 weeks), and peer-supported practice are associated with larger effects [3].

Two complementary strands of evidence speak to “why” AI-mediated instruction works. First, randomized and quasi-experimental studies with ASR show that immediate, objective feedback plus iterative practice produces measurable gains in accentedness, comprehensibility, and global speaking ability relative to traditional feedback [6]. Second, adding collaborative elements (peer correction around ASR transcripts) amplifies effects, precisely the moderator pattern found in the meta-analysis [3]. In writing, automated feedback similarly provides frequent, criterion-referenced guidance that accumulates into performance gains [11].

AI-supported activities affect the motivational “engine” of learning. In a controlled trial of AI-mediated instruction, students improved in L2 achievement, motivation and self-regulated learning relative to controls [12]. AI-augmented mobile/online practice lifts willingness to communicate (WTC) and class engagement at post-test; in one study, between-group differences favored the AI condition [7]. These results suggest that part of the performance benefit is mediated by increased practice quantity/quality and heightened motivation during technology-rich tasks [12,7].

The synthesized moderators help explain the spread in effects across studies and skills. Explicit, informative feedback outperforms purely implicit signals in ASR-based pronunciation learning [3]. Segmental targets show stronger gains than suprasegmentals; evidence for intelligibility/comprehensibility measures is emerging but thinner [3,4]. Medium to long interventions (≈6-24 weeks) tend to deliver larger, more stable improvements than brief exposures [3]. Studies with peer practice/support show notably larger effects than purely individual practice [3]; this aligns with RCT evidence where ASR combined with peer correction outperformed teacher-led feedback alone [6]. In learner profile, effects are often larger for adults and intermediate-proficiency learners [3].

Results from an intelligent tutoring system for reading comprehension (not language production) show that well-designed, feedback-rich, web-based tutoring can produce standardized gains on external tests with moderate, targeted improvements on proximal measures, even with relatively light weekly dosage [13].

In both ASR and AWE, designs that diagnose specific issues and point learners toward concrete fixes outperform generic signals [3,11]. Leverage well-validated apps but monitor classroom transfer; an AI-powered mobile app improved fluency/pronunciation in controlled classroom implementations [14], yet systematic reviews caution that much ASR research is quasi-experimental and segmental-focused; classroom translation and broader outcome coverage need attention [14,4].

Limitations of the literature and of our synthesis Despite promising effects, several constraints deserve emphasis. First, ASR studies emphasize segmental accuracy and short-term outcomes; evidence for suprasegmentals, intelligibility, and longer-term transfer is comparatively sparse. Second, non-performance outcomes (motivation, WTC, self-regulation) are frequently self-reported and unblinded, increasing measurement bias risk. Third, the modality mixes in trials (adjusted standardized effects vs. post-test SMDs) and cross-skill pooling introduce heterogeneity; separating domain-specific pools as evidence accumulates will sharpen generalizability.

CONCLUSION

This meta-analysis show that artificial intelligence-assisted language learning significantly improves second-language performance, with good benefits in writing, speaking, listening, and reading domains. The strongest improvements were observed in listening and speaking, suggesting that AI tools provide effective, feedback-driven practice in communicative tasks. Moderate effects in writing and reading further highlight the potential of AI to support diverse language skills. Narrative findings indicate reduced anxiety and improved learner engagement.

References

1) Nazari N, Shabbir MS, Setiawan R. Application of Artificial Intelligence powered digital writing assistant in higher education: randomized controlled trial. *Heliyon.* 2021;7(4): e07014. doi: 10.1016/j.heliyon.2021.e07014.

2) Wei P, Wang X, Dong H. The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial. *Front Psychol.* 2023; 14:1249991. doi:10.3389/fpsyg.2023.1249991.

3) Ngo TTN, Chen HHJ, Lai KK. The effectiveness of automatic speech recognition in ESL/EFL pronunciation: a meta-analysis. *ReCALL.* 2024;36(1):4-21. doi:10.1017/S0958344023000113.

4) Liu Y, Ab Rahman F, Mohamad Zain F. A systematic literature review of research on automatic speech recognition in EFL pronunciation. *Cogent Educ.* 2025;12(1):2466288. doi:10.1080/2331186X.2025.2466288.

5) Xiao Y. The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: a randomized controlled trial. *Humanit Soc Sci Commun.* 2025; 12:425. doi:10.1057/s41599-025-04672-8.

6) Sun W. The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. *Front Psychol.* 2023; 14:1210187. doi:10.3389/fpsyg.2023.1210187.

7) Ouyang Z, Jiang Y, Liu H. The Effects of Duolingo, an AI-Integrated Technology, on EFL Learners' Willingness to Communicate and Engagement in Online Classes. *Int Rev Res Open Distrib Learn (IRRODL).* 2024;25(3):97-115. doi:10.19173/irrodl. v25i3.7677.

8) Qiao H, Zhao A. Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Front Psychol.* 2023; 14:1255594. doi:10.3389/fpsyg.2023.1255594.

9) Wijekumar K, Meyer BJF, Lei P-W, Lin Y-C, Johnson LA, Spielvogel J, et al. High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders' content area reading comprehension. *Comput Educ.* 2013; 68:366-79. doi: 10.1016/j.compedu.2013.05.021.

10) Wu X-Y. Artificial intelligence in L2 learning: a meta-analysis of contextual, instructional, and social-emotional moderators. *System.* 2024; 126:103498.

11) Fleckenstein J, Liebenow LW, Meyer J. Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. *Front Artif Intell.* 2023; 6:1162454. doi:10.3389/frai.2023.1162454.

12) Wei L. Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Front Psychol.* 2023; 14:1261955. doi:10.3389/fpsyg.2023.1261955.

13) Wijekumar K, Meyer BJF, Lei P-W, Lin Y-C, Johnson LA, Spielvogel J, et al. Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. *J Res Educ Eff.* 2014;7(4):331-57. doi:10.1080/19345747.2013.853333.

14) Ma M, Noordin N, Razali AB. Improving EFL speaking performance among undergraduate students with an AI-powered mobile app in after-class assignments: an empirical investigation. *Humanit Soc Sci Commun.* 2025; 12:370.