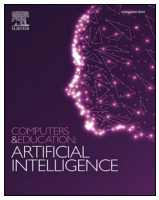
Computers and Education: Artificial Intelligence 9 (2025) 100522

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



# Effectiveness of Artificial Intelligence (AI) in language teaching

Peter Joseph Torres *, Yunus Emre Kahveci

*Department of English, Arizona State University, Tempe, AZ, USA*

ARTICLE INFO

*Keywords:*
Artificial intelligence
Education
English as a foreign language
Meta-analysis
Online language learning
Technology integration

ABSTRACT

This study examines the effectiveness of artificial intelligence (AI) in language teaching, particularly in English as a Foreign Language (EFL) classrooms, following AI's increased adoption after the COVID-19 pandemic. Through a multilevel meta-analysis of 117 effect sizes across 46 empirical studies published between 2022 and 2025, results show that AI has a statistically significant medium-to-large overall impact on language learning ($g = 0.74$, 95 % CI [0.57, 0.92], $p < .001$) across all five major skills, with vocabulary showing the strongest effects, followed by reading, writing, listening, and speaking.

Grounded in constructivist, adaptive learning, and cognitive load theories, moderator analyses revealed several key insights: (1) AI is more effective in face-to-face and blended settings than in fully online classrooms; (2) AI is particularly effective for younger K-12 learners, suggesting tools are pedagogically optimized for foundational language learning; (3) similar effectiveness outcomes across AI platforms suggest implementation matters more than the tool itself; (4) AI can facilitate task completion but not develop long-term autonomous learning habits like self-regulation; and (5) AI works best as a supplement rather than replacement for traditional teaching.

The results, when interpreted through the novelty effect, cognitive load theory, and attention economy frameworks, suggest a technology saturation effect: the failure of AI tools to capture distinctive attention, reduce cognitive burden, or secure focused engagement in an already technology-rich environment. The synthesis outlines the current state of EFL literature, which has been focused on Asia and the Middle East, offering practical insights for educators considering AI integration.

## 1. Introduction

The transition to remote instruction during the 2020–2022 COVID-19 pandemic catalyzed unprecedented technological adoption among language educators, with many continuing to use digital tools even after returning to face-to-face instruction. Investigations of artificial intelligence's (AI) applications in language teaching have centered on isolated language skills. This study addresses this gap by employing a multilevel meta-analytic approach that examines AI's effectiveness across various language skills (reading, writing, speaking, listening, and vocabulary) and contextual factors (educational level, instructional mode, and tool type, among others), with particular emphasis on English as a Foreign Language (EFL) classroom. Grounded constructivist (Piaget, 1972), adaptive learning (Zhang & Dong, 2024), and cognitive load theories (Sweller, 1988), this study analyzes a corpus of 46 empirical studies with 117 effect sizes (individual tests) published after 2022, when AI adoption significantly accelerated following the pandemic, to address the following research questions.

1. Is AI an effective tool in teaching English as a foreign language (EFL)?
2. Does AI's effectiveness vary under different conditions, such as the choice of AI tool, geography, grade level, study duration, and the mode of instruction, among other factors?

The results indicate that AI tools have a statistically significant, medium-to-large overall effect on language learning ($g = 0.74$, 95 % CI [0.57, 0.92], $p < .001$), with effectiveness varying substantially across different skills and contexts. Many moderators had a significant effect, suggesting that AI's effectiveness in language learning varies across contexts.

* Corresponding author.
*E-mail addresses:* p.torres@asu.edu (P.J. Torres), ykahveci@asu.edu (Y.E. Kahveci).

https://doi.org/10.1016/j.caeai.2025.100522
Received 18 September 2025; Received in revised form 30 November 2025; Accepted 30 November 2025
Available online 1 December 2025

## 2. Literature review

### 2.1. Language learning and AI after COVID-19

The integration of technology in language education has evolved dramatically since the COVID-19 pandemic. While Computer-Assisted Language Learning (CALL) had been established for decades, the pandemic catalyzed unprecedented technological adoption among language educators. Ahn and Chi (2023) found that the sudden move to digital platforms led language teachers to rapidly develop new technological competencies and that, despite initial struggles, many stayed open to using digital tools even after resuming in-person classes. As Moorhouse et al. (2023) observed, the post-pandemic period has witnessed rapidly increasing adoption of emerging technologies, particularly Generative AI, including chatbots, intelligent tutoring systems, and writing support tools in language education contexts.

While pre-pandemic studies investigated AI applications in language teaching, the current landscape reflects a fundamentally changed environment with more technology-receptive learning communities and increasingly sophisticated AI capabilities. Despite this rapid evolution, comprehensive syntheses of AI's effectiveness have typically focused on isolated language skills. This meta-analysis addresses this gap by examining AI effectiveness across all language-learning skills in EFL classrooms since 2022, when adoption accelerated significantly following the pandemic.

### 2.2. AI innovation in pedagogy and language education

Large Language Models (LLMs), as popularly known today, are AI systems based on transformer neural network architectures, trained on vast amounts of text data, that use deep learning to perform natural language processing and generate human-like text responses based on the input they receive. In educational contexts, these LLMs are often deployed as chatbots, or interactive applications designed for sustained conversational exchanges with learners (Huang et al., 2023). As Chen et al. (2022) document in their review of two decades of research on AI applications in education (AIED), AI's evolution has enabled truly personalized learning at scale. A quick survey of contemporary AI education tools trained in LLMs reveals their ability to ask, analyze, answer, classify, detect plagiarism, evaluate interlocutor knowledge, explain, generate content and feedback, measure semantic similarity and tone of writing, and offer 24/7 language practice opportunities beyond temporal and geographic barriers, among many others. Contemporary AI has fundamentally shifted the teacher's role from primary knowledge transmitter to learning facilitator who guides students' AI-enhanced practice. While AI encompasses diverse technical mechanisms such as neural networks, machine learning algorithms, and natural language processing (Chen et al., 2022), this study focuses on the functional application and pedagogical outcomes of AI tools in language learning contexts.

The pedagogical value of AI in language teaching is grounded in multiple learning theories. Constructivist theory (Piaget, 1972) emphasizes active knowledge construction, which Song and Song (2024) observed when AI tools enhanced the writing skills of Chinese students in blended EFL classrooms. AI in education encompasses various roles, including intelligent tutoring and learning partner systems that can provide personalized guidance and feedback (Hwang et al., 2020), which are particularly useful for language learning. Adaptive Learning theory, which advocates for dynamically tailored learning experiences (Zhang & Dong, 2024), aligns with AI's ability to provide individualized feedback and instruction that has been shown to improve EFL learners' writing (Liu et al., 2024), speaking (Qiao & Zhao, 2023), and vocabulary skills (Hsu et al., 2024). Similarly, cognitive load theory suggests that learning is optimized when unnecessary mental burdens are reduced. Kohnke (2024) demonstrated this principle when English for Academic Purposes students reported that AI tools decreased extraneous cognitive load through immediate, clear feedback. These theoretical frameworks collectively provide a foundation for understanding how AI might support language learning across different skill domains.

### 2.3. Meta-analysis on AI use in language learning

Since the rapid adoption of AI in education in 2022, research examining its effectiveness across various language-learning domains has increased. Regarding vocabulary acquisition, Liu and Chen (2023) demonstrated that using AI applications improved vocabulary acquisition in Taiwanese elementary schools. In reading comprehension, Kim's (2024) investigation found that ChatGPT helped EFL learners identify main ideas in a Korean high school. For speaking skills, Fathi et al. (2024) found that engaging with the Andy English chatbot improved EFL learners' fluency, coherence, and pronunciation at an Iranian university (See also Hwang et al., 2024 on mobile-assisted language learning). In terms of writing, Biju et al. (2024) found that AI's immediate feedback capabilities reduce student anxiety in Bangladeshi universities. As for listening skills, Li and Peng's (2022) study of an AI-based platform in Chinese EFL classrooms found no significant improvement in listening scores, yet it boosted student engagement. These examples preview the diverse AI research recently emerging in language teaching, underscoring the need for a comprehensive meta-analysis to capture AI's effect on language learning outcomes.

Meta-analysis is a systematic tool for assessing the current state of the field, using quantitative data from the primary studies to calculate effect sizes—the strength of relationships between outcomes and variables (Chong & Plonsky, 2021). Meta-analysis also examines moderators—factors that affect the resulting effect sizes—to determine whether the results arise from differences in study conditions. Some moderators regarding chatbots identified in the studies by Lee and Hwang (2022) and Zhang et al. (2023) include learners' characteristics (age, education level, language proficiency) and study design elements (duration, sample size, instructional modality). This meta-analysis makes a unique contribution to the discussion by examining AI's effectiveness across five language skills and a comprehensive list of moderators, rather than focusing on a particular skill or a few moderators. Recognizing these moderators is essential as it may inform EFL instructors' teaching strategies.

## 3. Methodology

### 3.1. Inclusion criteria

Articles were retrieved using an institutional library search engine that retrieves vetted, academically accessible research from various platforms: ERIC, JSTOR, Scopus, SAGE, Web of Science, Google Scholar, and EBSCO. Using the tool minimizes potential human error when conducting searches across different platforms and ensures organization, efficiency, and full-text access to all generated articles. Studies were then assessed using the inclusion/exclusion criteria presented in Table 1, and the process is outlined in Fig. 1. The study was guided by the screening conventions set by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The list of articles is found in Appendix A.

### 3.2. Multilevel analysis

As the corpus consists of studies featuring multiple experiments and, thus, multiple effect sizes and the likelihood of high heterogeneity, a multilevel (random-effects) meta-analysis is the appropriate tool, as it accounts for between-study variability and addresses potential dependencies (Van den Noortgate et al., 2013). As noted by Cooper et al. (2009), preserving all effect sizes enhances statistical power, improves our understanding of the intervention, and captures the subtle variations in outcomes (See Appendix B for complete data). By including diverse

**Table 1**
Inclusion/exclusion criteria.

| Inclusion criteria | Exclusion criteria |
|---|---|
| Empirical studies on language learning. | Non-empirical studies (literature syntheses, theoretical papers, AI reviews, etc.) |
| Peer-reviewed studies published from 2022 onwards | Non-peer-reviewed studies (conference proceedings, preprints, abstracts) or pre-2022 studies |
| Studies investigating GenAI's effects on language skills in EFL classrooms | Studies not investigating Gen AI's effect on language learning |
| Studies providing adequate statistical information: number of participants, mean, and standard deviation for both treatment and control groups | Studies not providing sufficient statistical data for effect-size calculation |
| Studies conducted in a classroom setting (face-to-face, hybrid, or online) | Studies conducted outside instructional settings (technology demonstrations, tool-development studies, etc.) |

effect sizes, everything is accounted for and reported (Borenstein et al., 2011).

The rest of the analyses are canon to meta-analysis literature, including initial, outlier, sensitivity, and moderator analyses, heterogeneity measures, and publication bias. All analyses were conducted using R (version 4.5.1) with the metafor package (version 4.8.0; Viechtbauer, 2010). We used the standardized mean difference (SMD) as the effect size metric to synthesize the results across studies. Effect sizes were calculated as Hedges' *g*, which is Cohen's *d* corrected for small-sample bias, making it suitable for meta-analysis. (Cohen, 2013; Hedges & Olkin, 2014). Cook's distance identifies influential cases by measuring how individual effect sizes impact overall model parameters. Outliers were identified using the conservative 4/n threshold (where n = number of effect sizes), which adjusts for sample size while avoiding excessive data removal (Viechtbauer & Cheung, 2010).

### 3.3. Variables

#### 3.3.1. Performance and affective skills

Expanding on Plonsky and Oswald's (2014) call for inclusivity in selecting articles for meta-analysis, this study incorporates both direct performance indicators (e.g., fluency, accuracy, and pronunciation) and affective variables (e.g., willingness to communicate, motivation, and self-efficacy). This framework acknowledges that language proficiency involves linguistic skills and psychological factors affecting acquisition (Dornyei & Ryan, 2015). Larsen-Freeman (2015) states that language learning comprises complex systems with interacting components. Thus, focusing solely on performance measures without accounting for psychological factors may lead to an incomplete understanding of intervention effectiveness.

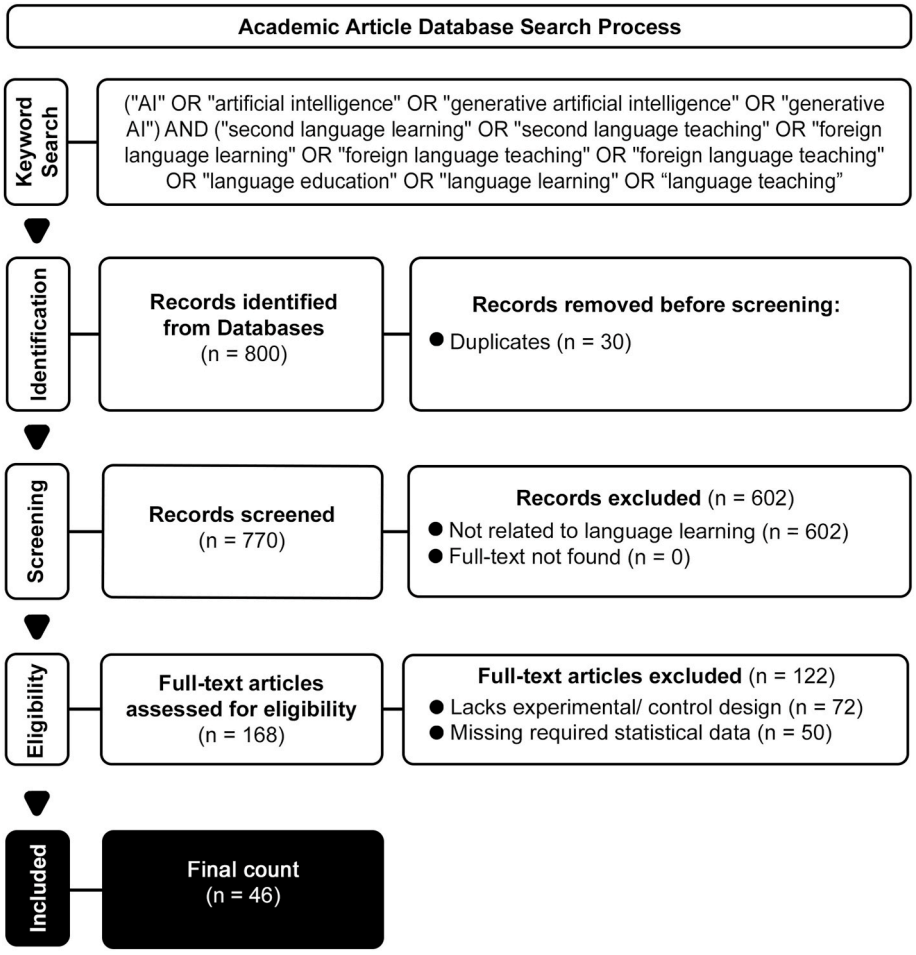



**Fig. 1.** Article selection criteria.
PRISMA diagram illustrating the article screening process for the meta-analysis.

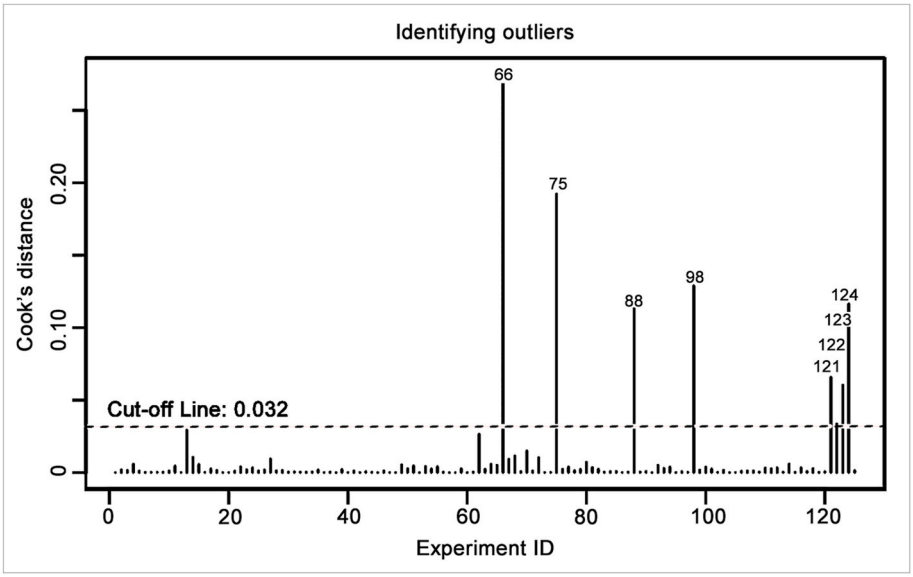



**Fig. 2.** Cook's distance for identifying outliers.
Cook's distance values identified eight outliers using the calculated threshold of 0.032 (horizontal line). The adjusted model retains 117 effect sizes after outlier removal.

**Table 2**
Comparison of Initial (with outliers) and Adjusted Models (without outliers).

| | *k* | Hedges' *g* | 95 % CI | $I^2$ (%) | $\tau^2$ | SE | AIC |
|---|---|---|---|---|---|---|---|
| Initial Model | 125 | 1.17*** | [0.75, 1.60] | 97.23 | 1.80 | 0.22 | 870 |
| Adjusted Model | 117 | 0.74*** | [0.57, 0.92] | 87.90 | 0.35 | 0.09 | 679 |

Note: *p < .05, **p < .01, ***p < .001.
*k* = number of effect sizes; Hedges'*g* = effect size; CI = confidence interval; $I^2$ = percentage of variance attributable to true heterogeneity; $\tau^2$ = between-study variance; SE = standard error; AIC = Akaike Information Criterion (lower values indicate better model fit).

### 3.3.2. *Moderator variables*

Moderator analysis involved systematically extracting and categorizing study design elements from the analyzed research papers. Rather than maintaining numerous similar variables as separate categories, we combined conceptually related or synonymous moderators into unified categories to ensure appropriate statistical representation and more meaningful interpretation of differential effects (e.g., the category "motivation" includes both "motivation" and "willingness"). Data extraction followed the coding-in-tandem method (Torres, 2021), in which both researchers convened virtually to categorize and address discrepancies in real time, thereby establishing consistent consolidation procedures. The clear, straightforward nature of the data facilitated a smooth process and quick agreement between researchers. A detailed table outlining each moderator category and its elements, using the original terminology from the studies, is provided in Appendix C.

For this meta-analysis, AI tools were categorized based on their implementation in primary studies rather than their underlying technical architecture. This approach is justified for several reasons: (1)

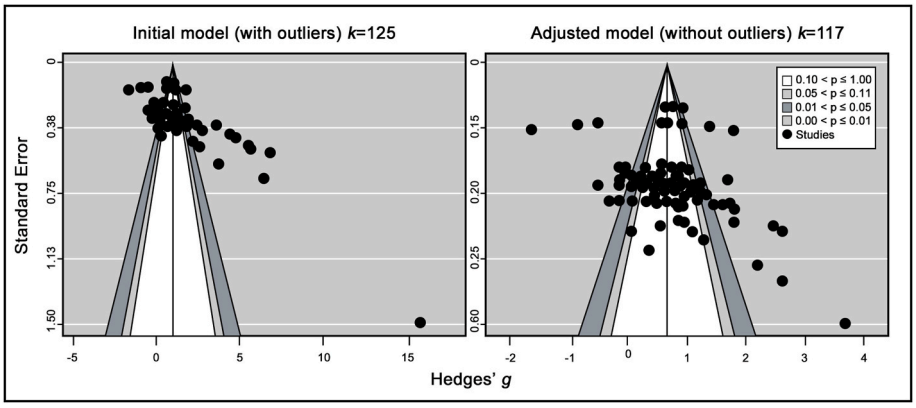



**Fig. 3.** Funnel plots for assessing Publication Bias.
The initial model (left) shows asymmetric study distribution, suggesting publication bias, while the adjusted model (right) displays greater symmetry around the central axis, indicating reduced bias after outlier removal.

primary studies often lack detailed technical specifications; (2) from a pedagogical perspective, the way tools are implemented may hold more significance than the continuously evolving algorithms; (3) many tools integrate multiple AI approaches; and (4) this study aims to inform educational practice, where tools are evaluated based on user-facing features.

## 4. Results

### 4.1. Initial analysis

The initial multilevel model yielded a significant positive effect (Hedges' $g = 1.17$, 95 % CI [0.75, 1.60], $p < .0001$) but the substantial heterogeneity ($Q(124) = 1930.22$, $p < .0001$, $I^2 = 97.23$ %, $\tau^2 = 1.8$, SE = 0.24), warrants further outlier and moderator analyses.

### 4.2. Outlier and sensitivity analysis

Cook's distance values identified eight outliers that exceeded the calculated threshold (Fig. 2). Table 2 presents model parameters before and after outlier removal.

Although both models are significant ($p < .001$) with large effects, the adjusted model demonstrated improved precision (37 % reduction in effect size, 59 % reduction in standard error, 22 % decrease in AIC) and greater consistency across studies (81 % reduction in between-study variance, $\tau^2$), indicating that the initial estimates were inflated by extreme values.

### 4.3. Publication bias

Statistical tests for publication bias become unreliable when heterogeneity is substantial, as the tests cannot distinguish between actual bias and genuine study differences (Afonso et al., 2024; Ioannidis & Trikalinos, 2007). This challenge is common in meta-analyses of AI tools in language education where skills vary immensely (e.g., Guan et al., 2024; Wu & Yu, 2024). Therefore, this study employs visual funnel plot inspection shown in Fig. 3 (Sterne et al., 2011).

The improved symmetry after removing outliers suggests heterogeneity reflects genuine diversity in interventions and contexts rather than systematic publication bias.

### 4.4. Moderator analysis

To explain the high heterogeneity, we examined categorical moderators that may account for differential AI effectiveness across studies, as shown in Table 3.

All language skills demonstrated significant positive effects, with vocabulary showing the strongest effect and speaking the weakest. All five skills represent medium-to-large effects by conventional standards. We did not restrict our inclusion criteria to any specific skill based on data availability, allowing us to demonstrate where recent research has concentrated.

Table 4 examines moderators that may explain AI's differential effectiveness across skills. A list of what each specific skill entails, based on the primary studies, can be found in Appendix C. Comprehension (reading) demonstrated the strongest effect, representing an exceptionally large impact, though the limited number of studies affects the interpretability of this result. Among well-represented skills, overall speaking and self-efficacy demonstrated substantial effects. Notably, self-regulation had no substantial effect, and accuracy did not reach significance. Geographically, the Middle East showed stronger effects compared to East Asia. Online-only interventions failed to reach statistical significance, while both in-person and blended approaches showed significant effects. AI interventions showed stronger effects among K-12 students than among college students. ChatGPT and other chatbots demonstrated significant effectiveness, while Duolingo showed large but non-significant effects.

Table 5 shows significant QM values (all $p < .001$), indicating that differences across subcategories within each moderator category are meaningful and not due to chance. Specific skills (e.g., fluency and accuracy) showed the strongest moderating effect (QM = 240.32), substantially exceeding main skills (e.g., vocabulary and speaking) (QM = 70.92), suggesting fine-grained skill distinctions are more predictive than broad categories. However, QE values, which test whether variation remains after controlling for moderators, remain significant ($p < .001$), indicating that unmeasured factors continue to affect AI effectiveness.

**Table 3**
AI's effectiveness by skill (without outliers).

| | *k* | *g* | SE | 95 % CI | *z* | *p* |
|---|---|---|---|---|---|---|
| Listening | 10 | 0.65 | 0.19 | [0.28,1.02] | 3.43 | <0.001 |
| Reading | 18 | 0.84 | 0.22 | [0.41,1.27] | 3.84 | <0.001 |
| Speaking | 35 | 0.62 | 0.18 | [0.27,0.98] | 3.49 | <0.001 |
| Vocabulary | 11 | 0.93 | 0.24 | [0.46,1.41] | 3.84 | <0.001 |
| Writing | 43 | 0.73 | 0.14 | [0.46,1.00] | 5.33 | <0.001 |

Note: *p < .05, **p < .01, ***p < .001.
*k* = number of effect sizes; *g* = Hedges' *g*; SE = Standard Error; 95 % CI = confidence interval; *z* = test statistic for individual effect sizes; *p* = significance level.

**Table 4**
Moderator analysis (without outliers).

| Moderator | *k* | *g* | SE | 95 % CI | *z* | *p* |
|---|---|---|---|---|---|---|
| **Specific Skill** | **117** | | | | | |
| Accuracy | 10 | 0.27 | 0.15 | [-0.02,0.56] | 1.83 | 0.068 |
| Fluency | 9 | 0.63 | 0.17 | [0.03,0.95] | 3.75 | <0.001*** |
| Listening | 8 | 0.82 | 0.19 | [0.44,1.19] | 4.29 | <0.001*** |
| Motivation | 13 | 0.82 | 0.14 | [0.54,1.10] | 5.81 | <0.001*** |
| Organization and Development | 12 | 0.43 | 0.16 | [0.11,0.75] | 2.63 | <0.01** |
| Overall speaking | 12 | 1.18 | 0.15 | [0.88,1.48] | 7.76 | <0.001*** |
| Overall writing | 13 | 0.53 | 0.14 | [0.26,0.81] | 3.77 | <0.001*** |
| Comprehension | 5 | 1.97 | 0.19 | [1.61,2.34] | 10.64 | <0.001*** |
| Self-efficacy | 10 | 0.98 | 0.14 | [0.70,1.26] | 6.85 | <0.001*** |
| Self-regulation | 9 | 0.01 | 0.14 | [-0.26,0.28] | 0.04 | 0.968 |
| Vocabulary Knowledge | 16 | 0.60 | 0.16 | [0.28,0.91] | 3.74 | <0.001*** |
| **Education Level** | **117** | | | | | |
| College | 95 | 0.70 | 0.10 | [0.51,0.90] | 7.14 | <0.001*** |
| K-12 | 22 | 0.90 | 0.19 | [0.52,1.28] | 4.69 | <0.001*** |
| Primary | 16 | 1.04 | 0.26 | [0.54,1.54] | 4.06 | <0.001*** |
| Secondary | 6 | 0.72 | 0.29 | [0.15,1.29] | 2.46 | 0.014* |
| **Mode** | **117** | | | | | |
| Blended | 44 | 0.82 | 0.14 | [0.55,1.09] | 5.88 | <0.001*** |
| In-person | 63 | 0.73 | 0.12 | [0.50,0.96] | 6.16 | <0.001*** |
| Online | 10 | 0.28 | 0.39 | [-0.48,1.05] | 0.73 | 0.463 |
| **Duration** | **117** | | | | | |
| Long | 47 | 0.81 | 0.14 | [0.54,1.09] | 5.82 | <0.001*** |
| Short | 70 | 0.70 | 0.11 | [0.47,0.92] | 6.11 | <0.001*** |
| **AI Tool** | **117** | | | | | |
| Chatbot | 22 | 0.84 | 0.24 | [0.37,1.3] | 3.53 | <0.001*** |
| ChatGPT | 26 | 0.81 | 0.17 | [0.47,1.15] | 4.68 | <0.001*** |
| Duolingo | 5 | 0.95 | 0.56 | [-0.14,2.04] | 1.71 | 0.087 |
| Other | 64 | 0.68 | 0.12 | [0.44,0.92] | 5.60 | <0.001*** |
| **Geographic Area** | **117** | | | | | |
| East Asia | 79 | 0.65 | 0.11 | [0.43,0.87] | 5.85 | <0.001*** |
| Middle East | 33 | 0.92 | 0.15 | [0.62,1.22] | 6.04 | <0.001*** |
| Other | 5 | 0.71 | 0.33 | [0.08,1.35] | 2.19 | 0.028* |

Note: *p < .05, **p < .01, ***p < .001.

**Table 5**
Summary of heterogeneity tests (QE) and moderator effects (QM).

| Moderator Categories | *k* | QE | df | *p* | QM | df | *p* |
|---|---|---|---|---|---|---|---|
| Main Skill | 117 | 1018.07 | 112 | <0.001*** | 70.92 | 5 | <0.001*** |
| Specific Skill | 117 | 948.31 | 106 | <0.001*** | 240.32 | 11 | <0.001*** |
| Education Level | 117 | 1064.92 | 115 | <0.001*** | 72.94 | 2 | <0.001*** |
| Mode of Instruction | 117 | 1058.42 | 114 | <0.001*** | 73.11 | 3 | <0.001*** |
| Duration | 117 | 1076.50 | 115 | <0.001*** | 71.24 | 2 | <0.001*** |
| AI Tool | 117 | 1051.96 | 113 | <0.001*** | 68.69 | 4 | <0.001*** |
| Geographic Area | 117 | 1079.96 | 114 | <0.001*** | 73.87 | 3 | <0.001*** |

Note: *p < .05, **p < .01, ***p < .001.
QE = test for residual heterogeneity; QM = test of moderators; df = degrees of freedom.

## 5. Discussion

We interpret our findings through three core theoretical frameworks that guided this study's design: Constructivist Theory (Section 5.1), which predicts AI effectiveness varies by social context; Cognitive Load Theory (Sections 5.1, 5.2), which predicts differential effects by task complexity; and Adaptive Learning Theory (Section 5.5), which predicts that individualized scaffolding adapts differently across learner populations. To further contextualize our findings, we also draw on Social Cognitive Theory (Section 5.3), novelty effect, and attention economy frameworks (Section 5.2), which help explain unexpected patterns in the data.

### 5.1. AI effectiveness by skill and tool

AI demonstrated statistically significant positive effects on language learning, with the adjusted model showing a moderate to large effect. Among the five main language skills, vocabulary acquisition showed to benefit the most from AI, aligning with Khan et al.'s (2024) findings on AI effectiveness in EFL instruction. These consistent positive effects across multiple AI platforms suggest that effectiveness depends more on implementation approach than on specific tools and their technical features.

However, not all outcomes showed consistent effects. Duolingo exhibited large but non-significant effects, though the small sample size limits interpretability. Similarly, AI showed non-significant effects on accuracy, suggesting that while AI may facilitate task performance during use, this does not necessarily transfer to underlying competencies assessed independently.

Theoretical Implications: These patterns align with our core theoretical frameworks. Constructivist theory (Piaget, 1972) predicts that learning occurs through active knowledge construction in social contexts, which helps explain why AI tools with communicative functions (e.g., chatbots, ChatGPT) demonstrated the strongest effects. Cognitive Load Theory (Sweller, 1988) further accounts for AI's success on objectively measurable skills such as vocabulary acquisition and reading comprehension, where immediate feedback reduces extraneous cognitive load. In contrast, complex skills like accuracy (e.g., correct word usage, appropriate tone) and revision require sustained practice with subjective, contextual feedback, making them harder to scaffold through automated systems alone.

Pedagogical Implications: The findings suggest AI tools demonstrate meaningful effectiveness for language learning. Given this baseline effectiveness across tools, implementation strategies and pedagogical design may matter more than specific tool selection. Educators should prioritize integration quality, focusing AI use on objectively measurable skills that benefit from instant feedback while maintaining active involvement when integrating AI for complex tasks requiring nuanced judgment.

### 5.2. Technology saturation effect

Our moderator analysis revealed a meaningful pattern: AI interventions showed significant effects in face-to-face and blended settings but not in online-only contexts. This pattern aligns with Cognitive Load Theory (Sweller, 1988), which suggests that in online contexts where learners already manage multiple digital tools, adding AI may increase cognitive load without sufficient benefit. Additional theoretical perspectives further contextualize this finding. Clark's (1985) novelty effect posits that introducing new technologies temporarily improves performance through increased interest, but this novelty inevitably fizzles out over time. In a post-COVID digital age, chatbots and apps are no longer as novel as they once were. Attention economy frameworks (Davenport & Beck, 2001; Simon, 1971) view attention as a finite resource in information-rich environments. Each digital platform, learning management system, and communication tool competes for mental engagement in online learning environments. AI may struggle to capture sufficient attention in technology-saturated contexts, but in face-to-face settings where digital technologies are less ubiquitous, AI occupies a clearer attentional niche.

Drawing on these theoretical foundations, we propose the technology saturation effect: a context-dependent phenomenon in which new educational technology, like AI, is introduced into already technology-rich environments, producing diminished learning benefits because multiple mechanisms simultaneously reduce pedagogical impact. The new tool neither captures distinctive attention (novelty), reduces cognitive burden (cognitive load), nor secures focused engagement (attention economy). This hypothesis emerged from our findings and aligned with the theoretical frameworks mentioned. We present it here to stimulate further theoretical inquiry, recognizing that meta-analysis serves dual functions of both hypothesis testing and generation, with novel hypotheses emerging particularly when investigations into sources of heterogeneity lead to further inquiry (Mikolajewicz & Komarova, 2019). Future studies could evaluate cognitive load across different levels of digital saturation while controlling for other pedagogical factors.

### 5.3. The self-regulation paradox

Self-regulation showing near-zero benefits from AI ($g = 0.01$) is noteworthy. Social Cognitive Theory emphasizes that self-regulation encompasses multiple behavioral skills (e.g., goal-setting, progress monitoring, behavior adjustment) that require extended practice to develop (Bandura, 1986; Zimmerman, 2013). This contrasts with other affective skills like self-efficacy and motivation, which can be triggered immediately by AI's novelty and successful interactions (Bognár & Khine, 2025).

Most self-regulation studies were short-term ($k = 6$ of 9 studies were ≤10 weeks), likely insufficient for self-regulatory habit formation. Our

finding aligns with Bognár and Khine (2025), who observed that self-regulation declined most in AI-enhanced classrooms because AI tools offer insufficient support for long-term self-regulatory development. AI tools can facilitate task completion without teaching self-regulation, creating a paradox where AI boosts confidence and interest in learning while failing to develop autonomous learning habits.

Pedagogical Implications: Results suggest that educators should not assume AI tools automatically foster self-regulation. Explicit instruction and scaffolding may be necessary alongside AI tools for skills requiring long-term practice or metacognitive control.

### 5.4. Geographic and educational context factors

Despite applying no geographical restrictions, every study fitting our inclusion criteria came from Asian and Middle Eastern contexts. The absence of studies from English-dominant regions is expected, as EFL research naturally occurs where English is taught as a foreign language. However, this regional concentration reveals a critical research gap: the lack of AI language education research from South America, sub-Saharan Africa, Eastern Europe, among others. These gaps reflect known disparities in resources, infrastructure, and cultural attitudes toward English learning and AI integration (see Canagarajah, 2002). As meta-analyses serve to catalog trends and identify gaps (Mikolajewicz & Komarova, 2019), we note these findings may not generalize beyond Asian and Middle Eastern EFL contexts.

Pedagogical Implications: Educators in underrepresented regions should pilot-test AI tools locally and document outcomes to help address this geographic research gap.

### 5.5. Effectiveness based on education level

AI interventions proved more effective with younger learners, showing stronger effects in primary education ($g = 1.04$) compared to college students ($g = 0.70$). Adaptive Learning Theory (Zhang & Dong, 2024) helps explain this pattern through its emphasis on how adaptive systems provide personalized scaffolding matched to learner needs. Zhang and Dong (2024) note that adaptive learning systems are particularly effective at supporting learners developing foundational competencies through individualized pacing and targeted feedback. Primary EFL students are precisely such learners, requiring structured support for discrete, measurable skills like vocabulary and grammar where current AI tools excel at providing immediate feedback and customized practice. University students, however, require support for complex, context-dependent competencies like disciplinary writing conventions, nuanced argumentation, and pragmatic discourse where current adaptive systems provide less targeted scaffolding. This pattern suggests current AI tools are pedagogically optimized for foundational language learning but require further development for advanced academic contexts.

Pedagogical Implications: Educators should recognize that current AI tools may be most effective for foundational language learning. University-level instruction may require more sophisticated implementation strategies addressing complex discourse, disciplinary writing, and critical analysis.

### 5.6. Limitations

First, to prioritize function and pedagogical outcomes, we aggregated AI tools into the product they represent (e.g., ChatGPT, Duolingo, chatbots) instead of technical specifications (e.g., transformer models, neural networks) because (1) primary studies rarely provide sufficient detail of the mechanism behind the tools they investigate, and (2) AI is marketed and adopted in educational contexts based more on what products can do than on their underlying architectures. Insufficient information contributes to the "other" tool category being highly heterogeneous, which reflects genuine diversity in AI implementation. While this approach potentially masks distinctions in technical mechanisms, it provides an accessible synthesis for educators.

Second, to accurately represent the current state of AI in language education, we included all qualifying studies, even when some moderators yielded small-sized moderator categories. Consequently, such moderators should be interpreted with caution.

Finally, although typical in applied linguistics and education research, this study acknowledges the high heterogeneity ($I^2 = 88$ %) of our model. Heterogeneity here reflects the genuine diversity of study outcomes ranging from vocabulary acquisition to organizational skills within a single meta-analysis framework (Afonso et al., 2024; Ioannidis & Trikalinos, 2007). Dong et al. (2025) reported comparable heterogeneity ($I^2 = 93$ %) in their meta-analysis of AI in education, demonstrating that random-effects modeling appropriately addresses heterogeneity by treating studies as having varying effects rather than assuming a single true effect. Our multilevel random-effects approach follows this same methodological standard for heterogeneous educational data. Practitioners would benefit from referring to the moderator analyses to determine which findings apply to their specific instructional contexts.

## 6. Conclusion

This multilevel meta-analysis synthesized 117 effect sizes from 46 empirical studies to examine AI's effectiveness in EFL classrooms after the increased AI use following the COVID pandemic (2022–2025). Our findings show that AI tools have significant positive impacts on language learning across all five major skills—vocabulary, reading, writing, listening, and speaking—with substantial variation across educational levels and instructional settings.

Drawing on Constructivist, Cognitive Load, and Adaptive Learning theories, moderator analyses revealed key patterns: First, instructional modality emerged as a significant factor, with AI demonstrating strong effects in face-to-face and blended settings but failing to produce significant benefits in online-only contexts. Second, AI interventions proved particularly effective for younger K-12 learners compared to university students, suggesting current tools are pedagogically optimized for foundational language learning but require further development for advanced academic contexts. Third, effectiveness remained relatively consistent across different AI platforms, indicating that implementation quality and pedagogical design matter more than specific tool selection. Fourth, AI can help with task completion and improve discrete skills like vocabulary, but it cannot develop independent learning habits or self-regulation. Fifth, educators should prioritize AI use for objectively measurable skills that benefit from instant feedback while maintaining active involvement when integrating AI for complex tasks requiring nuanced judgment.

Our findings, when interpreted through the novelty effect, cognitive load theory, and attention economy frameworks, suggest a technology saturation effect: the introduction of new AI tools neither captures distinctive attention, reduces cognitive burden, nor secures focused engagement in an already technology-rich environment. Future research should examine whether AI's diminished effectiveness in online contexts persists and what this means for implementation strategies.

## CRediT authorship contribution statement

**Peter Joseph Torres:** Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis. **Yunus Emre Kahveci:** Writing – review & editing, Validation, Data curation, Conceptualization.

## Statements on open data and ethics

As this study is a systematic literature review and does not involve human participants, the Institutional Review Board determined it to be

Not Human Research (IRB ID: STUDY00022181), and ethical approval is not required. The complete dataset with effect sizes, moderators, and coding schemes is in Appendices A, B, and C. Additional files available upon request to the author.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors would like to acknowledge Joe Buenker, M.S., of ASU Library, for his support with the corpus.

## Appendix A. Studies and relevant information

| ID | Article | ID | Article | ID | Article |
|---|---|---|---|---|---|
| ID01 | Borna et al. (2024) | ID17 | Tai and Chen (2024) | ID33 | Alrajhi (2025) |
| ID02 | Chen et al. (2022) | ID18 | Wang et al., (2024) | ID34 | Chen et al. (2025) |
| ID03 | Escalante et al. (2023) | ID19 | Namaziandost 2025) | ID35 | Nguyen (2025) |
| ID04 | Fathi et al. (2024) | ID20 | Yuan (2025) | ID36 | Abdellatif et al. (2024) |
| ID05 | Huang and Mizumoto (2024) | ID21 | Xiao (2025) | ID37 | Alazemi (2024) |
| ID06 | Jeon (2023) | ID22 | Rad (2025) | ID38 | Behforouz and Ghaithi (2024) |
| ID07 | Khan (2024) | ID23 | Shi and Shakibaei (2025) | ID39 | Tai and Chen (2024b) |
| ID08 | Kim (2024) | ID24 | Shaalan and Ahmad (2025) | ID40 | Oktarin et al. (2024) |
| ID09 | Liu and Chen (2022) | ID25 | Jamshed et al. (2025) | ID41 | Hwang et al. (2023) |
| ID10 | Li and Peng (2022) | ID26 | Jose (2025) | ID42 | Kim and Cha (2023) |
| ID11 | Liu et al. (2023) | ID27 | Pan et al. (2025) | ID43 | Sun (2023) |
| ID12 | Liu et al. (2024) | ID28 | Chen et al. (2025) | ID44 | Zahran (2025) |
| ID13 | Biju et al. (2024) | ID29 | Aladini et al. (2025) | ID45 | Elmaadaway et al. (2025) |
| ID14 | Qasem (2023) | ID30 | Allehyani et al. (2025) | ID46 | Mekheimer (2025) |
| ID15 | Qiao and Zhao (2023) | ID31 | Zheldibayeva (2025) | | |
| ID16 | Song and Song (2023) | ID32 | Apriani et al. (2025) | | |

## Appendix B. Complete Effect Size Data

| # | ID | Level | Country | Mode | Tool | Duration | Main Skill | Treatment group | | | Control group | | | Effect Size | | Cook's Distance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Primary Secondary Tertiary | | Classroom Blended In-person | | Short Long | Writing Speaking Vocabulary Reading Listening | n | Mean | SD | n | Mean | SD | g | Variance | |
| 1 | ID01 | T | Iran | I | Other | S | W | 32 | 81.26 | 7.97 | 32 | 70.51 | 11.49 | 1.07 | 0.07 | 0.000001 |
| 2 | ID01 | T | Iran | I | Other | S | W | 32 | 87.20 | 7.18 | 32 | 70.51 | 11.49 | 1.72 | 0.09 | 0.002014 |
| 3 | ID02 | P | Taiwan | I | Other | S | S | 30 | 57.50 | 17.57 | 26 | 48.54 | 20.68 | 0.46 | 0.07 | 0.001526 |
| 4 | ID03 | T | Asia Pacific | B | ChatGPT | S | W | 23 | 33.37 | 1.91 | 25 | 33.68 | 2.07 | −0.15 | 0.08 | 0.005881 |
| 5 | ID04 | T | Iran | B | Chatbot | L | S | 33 | 6.53 | 1.14 | 32 | 5.97 | 0.79 | 0.56 | 0.06 | 0.001082 |
| 6 | ID04 | T | Iran | B | Chatbot | L | S | 33 | 6.28 | 0.79 | 32 | 5.58 | 0.84 | 0.85 | 0.07 | 0.000203 |
| 7 | ID04 | T | Iran | B | Chatbot | L | S | 33 | 6.93 | 0.94 | 32 | 6.16 | 0.90 | 0.83 | 0.07 | 0.000246 |
| 8 | ID04 | T | Iran | B | Chatbot | L | S | 33 | 6.83 | 0.90 | 32 | 6.08 | 0.89 | 0.83 | 0.07 | 0.000243 |
| 9 | ID04 | T | Iran | B | Chatbot | L | S | 33 | 4.16 | 0.81 | 32 | 3.58 | 0.61 | 0.80 | 0.07 | 0.000310 |
| 10 | ID05 | T | Japan | I | ChatGPT | S | W | 35 | 3.79 | 1.14 | 45 | 3.17 | 1.12 | 0.54 | 0.05 | 0.001176 |
| 11 | ID05 | T | Japan | I | ChatGPT | S | W | 35 | 3.58 | 1.36 | 45 | 3.57 | 1.23 | 0.01 | 0.05 | 0.004649 |
| 12 | ID05 | T | Japan | I | ChatGPT | S | W | 35 | 4.52 | 0.87 | 45 | 3.68 | 1.01 | 0.87 | 0.06 | 0.000159 |
| 13 | ID06 | P | South Korea | I | Chatbot | S | V | 18 | 7.94 | 1.95 | 17 | 1.88 | 1.11 | 3.70 | 0.31 | 0.029486 |
| 14 | ID06 | P | South Korea | I | Chatbot | S | V | 18 | 5.89 | 1.78 | 17 | 1.88 | 1.11 | 2.62 | 0.21 | 0.010599 |
| 15 | ID06 | P | South Korea | I | Chatbot | S | V | 18 | 6.33 | 2.30 | 17 | 2.06 | 1.34 | 2.20 | 0.18 | 0.005622 |
| 16 | ID06 | P | South Korea | I | Chatbot | S | V | 18 | 4.28 | 1.96 | 17 | 2.06 | 1.34 | 1.28 | 0.14 | 0.000209 |
| 17 | ID07 | S | Afghanistan | I | ChatGPT | S | V | 30 | 10.57 | 3.40 | 30 | 5.47 | 2.00 | 1.80 | 0.09 | 0.002554 |
| 18 | ID07 | S | Afghanistan | I | ChatGPT | S | V | 30 | 8.90 | 3.43 | 30 | 7.13 | 3.95 | 0.47 | 0.07 | 0.001489 |
| 19 | ID08 | S | South Korea | I | ChatGPT | L | R | 20 | 3.15 | 0.92 | 20 | 2.30 | 0.78 | 0.98 | 0.11 | 0.000032 |
| 20 | ID09 | P | Taiwan | I | Other | L | V | 18 | 78.67 | 16.92 | 18 | 55.33 | 24.49 | 1.08 | 0.13 | 0.000002 |
| 21 | ID09 | P | Taiwan | I | Other | L | V | 18 | 34.67 | 16.21 | 18 | 28.00 | 3.50 | 0.56 | 0.12 | 0.001052 |
| 22 | ID10 | T | China | B | Other | L | L | 17 | 83.41 | 3.06 | 15 | 83.27 | 2.78 | 0.05 | 0.13 | 0.004010 |
| 23 | ID10 | T | China | B | Other | L | L | 11 | 73.64 | 7.72 | 16 | 70.19 | 10.43 | 0.35 | 0.16 | 0.001941 |
| 24 | ID11 | T | China | B | Other | S | W | 33 | 16.31 | 2.82 | 31 | 15.84 | 2.55 | 0.17 | 0.06 | 0.003318 |
| 25 | ID11 | T | China | B | Other | S | W | 33 | 23.03 | 3.26 | 31 | 20.68 | 5.67 | 0.51 | 0.06 | 0.001333 |
| 26 | ID11 | T | China | B | Other | S | W | 33 | 24.06 | 3.72 | 31 | 22.42 | 4.14 | 0.41 | 0.06 | 0.001806 |
| 27 | ID11 | T | China | B | Other | S | W | 33 | 24.48 | 4.69 | 31 | 27.29 | 6.53 | −0.49 | 0.06 | 0.009500 |
| 28 | ID12 | P | China | B | ChatGPT | S | W | 32 | 72.44 | 10.74 | 33 | 67.35 | 8.33 | 0.52 | 0.06 | 0.001250 |
| 29 | ID12 | P | China | B | ChatGPT | S | W | 32 | 3.57 | 0.49 | 33 | 3.33 | 0.52 | 0.47 | 0.06 | 0.001513 |
| 30 | ID12 | P | China | B | ChatGPT | S | W | 32 | 3.49 | 0.59 | 33 | 3.08 | 0.56 | 0.70 | 0.07 | 0.000561 |

(continued on next page)

(*continued*)

| # | ID | Level | Country | Mode | Tool | Duration | Main Skill | Treatment group | | | Control group | | | Effect Size | | Cook's Distance | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Primary Secondary Tertiary | | Classroom Blended In-person | | Short Long | Writing Speaking Vocabulary Reading Listening | n | Mean | SD | n | Mean | SD | g | Variance | | |
| 31 | ID12 | P | China | B | ChatGPT | S | W | 32 | 3.66 | 0.48 | 33 | 3.30 | 0.54 | 0.70 | 0.07 | 0.000589 | |
| 32 | ID12 | P | China | B | ChatGPT | S | W | 32 | 3.75 | 0.48 | 33 | 3.39 | 0.48 | 0.74 | 0.07 | 0.000454 | |
| 33 | ID12 | P | China | B | ChatGPT | S | W | 32 | 3.49 | 0.56 | 33 | 3.02 | 0.65 | 0.76 | 0.07 | 0.000391 | |
| 34 | ID12 | P | China | B | ChatGPT | S | W | 32 | 3.66 | 0.58 | 33 | 3.27 | 0.52 | 0.70 | 0.07 | 0.000575 | |
| 35 | ID13 | T | Bangladesh | I | ChatGPT | S | W | 35 | 11.68 | 6.40 | 35 | 9.51 | 4.36 | 0.39 | 0.06 | 0.001929 | |
| 36 | ID14 | T | Saudi Arabia | B | Chatbot | L | V | 20 | 14.85 | 3.63 | 20 | 11.70 | 3.21 | 0.90 | 0.11 | 0.000111 | |
| 37 | ID15 | T | China | B | Duolingo | L | S | 47 | 5.94 | 0.49 | 46 | 5.52 | 0.67 | 0.71 | 0.05 | 0.000553 | |
| 38 | ID15 | T | China | B | Duolingo | L | S | 47 | 5.78 | 0.67 | 46 | 5.09 | 0.63 | 1.05 | 0.05 | 0.000001 | |
| 39 | ID15 | T | China | B | Duolingo | L | S | 47 | 6.32 | 0.93 | 46 | 4.97 | 0.56 | 1.74 | 0.06 | 0.002197 | |
| 40 | ID15 | T | China | B | Duolingo | L | S | 47 | 5.91 | 0.77 | 46 | 5.13 | 0.92 | 0.91 | 0.05 | 0.000102 | |
| 41 | ID15 | T | China | B | Duolingo | L | S | 47 | 4.34 | 0.82 | 46 | 3.89 | 0.83 | 0.54 | 0.04 | 0.001200 | |
| 42 | ID16 | T | China | B | ChatGPT | L | W | 25 | 59.12 | 14.23 | 25 | 45.18 | 15.62 | 0.92 | 0.09 | 0.000091 | |
| 43 | ID16 | T | China | B | ChatGPT | L | W | 25 | 15.96 | 3.71 | 25 | 13.71 | 3.12 | 0.65 | 0.08 | 0.000740 | |
| 44 | ID16 | T | China | B | ChatGPT | L | W | 25 | 16.56 | 3.54 | 25 | 13.63 | 3.63 | 0.80 | 0.09 | 0.000288 | |
| 45 | ID16 | T | China | B | ChatGPT | L | W | 25 | 19.89 | 4.82 | 25 | 15.89 | 4.12 | 0.88 | 0.09 | 0.000148 | |
| 46 | ID16 | T | China | B | ChatGPT | L | W | 25 | 20.06 | 3.33 | 25 | 18.21 | 3.58 | 0.53 | 0.08 | 0.001215 | |
| 47 | ID17 | P | Taiwan | B | ChatGPT | S | S | 28 | 63.21 | 16.50 | 29 | 49.34 | 11.82 | 0.96 | 0.08 | 0.000050 | |
| 48 | ID17 | P | Taiwan | B | ChatGPT | S | S | 28 | 62.96 | 10.89 | 29 | 49.34 | 11.82 | 1.18 | 0.08 | 0.000062 | |
| 49 | ID18 | T | China | O | Chatbot | S | S | 33 | 5.61 | 2.03 | 33 | 5.76 | 1.35 | −0.09 | 0.06 | 0.005409 | |
| 50 | ID18 | T | China | O | Chatbot | S | S | 33 | 6.21 | 2.06 | 33 | 5.76 | 1.35 | 0.26 | 0.06 | 0.002753 | |
| 51 | ID18 | T | China | O | Chatbot | S | S | 33 | 4.01 | 1.08 | 33 | 4.02 | 0.78 | −0.01 | 0.06 | 0.004754 | |
| 52 | ID18 | T | China | O | Chatbot | S | S | 33 | 4.69 | 0.49 | 33 | 4.02 | 0.78 | 1.02 | 0.07 | 0.000010 | |
| 53 | ID18 | T | China | O | Chatbot | S | S | 33 | 3.36 | 1.18 | 33 | 3.30 | 0.91 | 0.06 | 0.06 | 0.004206 | |
| 54 | ID18 | T | China | O | Chatbot | S | S | 33 | 3.61 | 1.26 | 33 | 3.30 | 0.91 | 0.28 | 0.06 | 0.002601 | |
| 55 | ID18 | T | China | O | Chatbot | S | S | 33 | 3.49 | 1.02 | 33 | 3.43 | 0.68 | 0.07 | 0.06 | 0.004109 | |
| 56 | ID18 | T | China | O | Chatbot | S | S | 33 | 4.35 | 0.66 | 33 | 3.43 | 0.68 | 1.36 | 0.07 | 0.000395 | |
| 57 | ID19 | T | Iran | B | Other | S | W | 35 | 46.08 | 7.72 | 35 | 38.42 | 6.73 | 1.05 | 0.06 | 0.000001 | |
| 58 | ID19 | T | Iran | B | Other | S | W | 35 | 17.57 | 1.59 | 35 | 14.85 | 3.03 | 1.11 | 0.07 | 0.000011 | |
| 59 | ID20 | T | China | I | Other | L | R | 150 | 82.60 | 6.20 | 150 | 70.20 | 7.50 | 1.80 | 0.02 | 0.002724 | |
| 60 | ID20 | T | China | I | Other | L | R | 150 | 4.30 | 0.50 | 150 | 3.80 | 0.60 | 0.90 | 0.01 | 0.000120 | |
| 61 | ID20 | T | China | I | Other | L | R | 150 | 2.50 | 0.60 | 150 | 3.40 | 0.70 | 1.38 | 0.02 | 0.000482 | * |
| 62 | ID20 | T | China | I | Other | L | R | 150 | 3.10 | 0.50 | 150 | 4.00 | 0.60 | −1.63 | 0.02 | 0.026644 | |
| 63 | ID21 | T | China | B | Other | S | L | 42 | 4.23 | 0.36 | 42 | 4.11 | 0.40 | 0.31 | 0.05 | 0.002422 | |
| 64 | ID21 | T | China | B | Other | S | L | 42 | 2.89 | 0.63 | 42 | 2.97 | 0.57 | −0.13 | 0.05 | 0.005905 | |
| 65 | ID21 | T | China | B | Other | S | L | 42 | 3.72 | 0.55 | 42 | 3.69 | 0.61 | −0.05 | 0.05 | 0.005172 | * |
| 66 | ID22 | T | Iran | B | Other | L | R | 30 | 28.21 | 0.99 | 30 | 13.11 | 0.89 | 15.83 | 2.16 | 0.275494 | † |
| 67 | ID22 | T | Iran | B | Other | L | R | 30 | 4.78 | 0.98 | 30 | 2.67 | 0.69 | 2.46 | 0.12 | 0.009272 | |
| 68 | ID22 | T | Iran | B | Other | L | R | 30 | 4.67 | 0.99 | 30 | 2.17 | 0.89 | 2.62 | 0.12 | 0.011642 | |
| 69 | ID23 | T | Iran | I | Other | L | S | 145 | 17.38 | 3.03 | 142 | 15.05 | 3.95 | 0.66 | 0.01 | 0.000742 | |
| 70 | ID23 | T | Iran | I | Other | L | S | 145 | 54.25 | 11.10 | 142 | 45.03 | 9.77 | −0.88 | 0.02 | 0.015035 | * |
| 71 | ID23 | T | Iran | I | Other | L | S | 145 | 59.68 | 14.01 | 142 | 52.65 | 10.58 | 0.56 | 0.01 | 0.001134 | |
| 72 | ID23 | T | Iran | I | Other | L | S | 145 | 65.16 | 17.04 | 142 | 57.71 | 10.97 | −0.52 | 0.01 | 0.010319 | * |
| 73 | ID24 | T | Saudi Arabia | I | Other | S | L | 31 | 23.68 | 1.88 | 33 | 19.67 | 4.13 | 1.22 | 0.07 | 0.000114 | |
| 74 | ID24 | T | Saudi Arabia | I | Other | S | S | 31 | 16.13 | 2.33 | 33 | 13.70 | 1.33 | 1.28 | 0.08 | 0.000206 | |
| 75 | ID25 | T | India | I | Chatbot | L | W | 56 | 39.05 | 0.87 | 56 | 32.37 | 1.06 | 6.84 | 0.24 | 0.192264 | † |
| 76 | ID26 | T | Oman | I | Other | S | R | 24 | 65.21 | 12.94 | 24 | 60.50 | 17.80 | 0.30 | 0.08 | 0.002422 | |
| 77 | ID26 | T | Oman | I | Other | S | R | 24 | 66.96 | 13.57 | 24 | 65.88 | 15.12 | 0.07 | 0.08 | 0.003972 | |
| 78 | ID26 | T | Oman | I | Other | S | R | 24 | 28.75 | 11.51 | 24 | 34.38 | 11.08 | 0.49 | 0.09 | 0.001377 | * |
| 79 | ID26 | T | Oman | I | Other | S | R | 24 | 23.17 | 20.88 | 24 | 34.42 | 44.59 | 0.32 | 0.08 | 0.002300 | * |
| 80 | ID26 | T | Oman | I | Other | S | R | 24 | 50.54 | 26.07 | 24 | 43.88 | 20.14 | −0.28 | 0.08 | 0.007103 | * |
| 81 | ID26 | T | Oman | I | Other | S | R | 24 | 3.21 | 2.45 | 24 | 3.54 | 2.59 | 0.13 | 0.08 | 0.003560 | * |
| 82 | ID26 | T | Oman | I | Other | S | R | 24 | 0.96 | 1.30 | 24 | 1.33 | 1.17 | 0.29 | 0.08 | 0.002443 | * |
| 83 | ID27 | T | China | B | Other | L | R | 31 | 3.75 | 0.29 | 30 | 3.50 | 0.36 | 0.76 | 0.07 | 0.000410 | |
| 84 | ID28 | T | China | I | Other | S | S | 29 | 3.18 | 0.56 | 24 | 2.80 | 0.65 | 0.62 | 0.08 | 0.000832 | |
| 85 | ID29 | T | Iran | I | Other | S | W | 276 | 25.80 | 8.70 | 285 | 20.81 | 7.26 | 0.62 | 0.01 | 0.000894 | |
| 86 | ID29 | T | Iran | I | Other | S | W | 276 | 18.91 | 3.90 | 285 | 14.95 | 4.49 | 0.94 | 0.01 | 0.000072 | |
| 87 | ID29 | T | Iran | I | Other | S | W | 276 | 52.78 | 4.31 | 285 | 48.95 | 5.70 | 0.76 | 0.01 | 0.000442 | |
| 88 | ID30 | T | Saudi Arabia | I | Other | S | R | 43 | 32.94 | 1.52 | 49 | 23.85 | 1.67 | 5.63 | 0.22 | 0.113235 | † |
| 89 | ID31 | T | Kazakhstan | I | Chatbot | L | L | 48 | 6.21 | 1.68 | 45 | 5.09 | 1.35 | 0.73 | 0.05 | 0.000506 | |
| 90 | ID31 | T | Kazakhstan | I | Chatbot | L | W | 48 | 3.62 | 0.73 | 45 | 3.18 | 0.65 | 0.63 | 0.05 | 0.000831 | |
| 91 | ID32 | T | Indonesia | I | ChatGPT | S | W | 25 | 81.11 | 22.98 | 25 | 60.30 | 20.18 | 0.95 | 0.09 | 0.000058 | |
| 92 | ID33 | T | Saudi Arabia | B | Other | S | V | 37 | 13.51 | 5.62 | 37 | 13.81 | 5.02 | −0.06 | 0.05 | 0.005177 | |
| 93 | ID34 | T | China | I | Other | S | W | 35 | 6.03 | 0.98 | 35 | 5.81 | 0.93 | 0.23 | 0.06 | 0.002949 | |
| 94 | ID34 | T | China | I | Other | S | W | 35 | 6.49 | 1.22 | 35 | 6.37 | 1.21 | 0.10 | 0.06 | 0.003895 | |
| 95 | ID35 | T | Vietnam | B | ChatGPT | L | W | 38 | 6.89 | 0.59 | 38 | 6.12 | 0.64 | 1.24 | 0.06 | 0.000141 | |
| 96 | ID36 | T | Saudi Arabia | I | Other | S | L | 30 | 17.38 | 1.60 | 27 | 13.83 | 3.07 | 1.45 | 0.09 | 0.000691 | |
| 97 | ID37 | T | Kuwait | I | Other | S | R | 40 | 17.75 | 2.30 | 40 | 16.00 | 2.00 | 0.80 | 0.05 | 0.000299 | |
| 98 | ID38 | T | Oman | I | Other | S | R | 30 | 18.46 | 0.97 | 30 | 9.30 | 1.74 | 6.42 | 0.41 | 0.128981 | † |
| 99 | ID39 | S | Taiwan | I | Chatbot | L | L | 31 | 91.58 | 22.47 | 31 | 81.35 | 25.25 | 0.42 | 0.07 | 0.001747 | |
| 100 | ID39 | S | Taiwan | I | Chatbot | L | L | 30 | 83.73 | 28.88 | 31 | 81.35 | 25.25 | 0.09 | 0.07 | 0.003945 | |

(*continued on next page*)

(*continued*)

| # | ID | Level | Country | Mode | Tool | Duration | Main Skill | Treatment group | | | Control group | | | Effect Size | | Cook's Distance | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Primary Secondary Tertiary | | Classroom Blended In-person | | Short Long | Writing Speaking Vocabulary Reading Listening | n | Mean | SD | n | Mean | SD | g | Variance | | |
| 101 | ID40 | T | Indonesia | B | ChatGPT | L | W | 25 | 86.48 | 3.18 | 25 | 77.28 | 6.34 | 1.81 | 0.11 | 0.002508 | |
| 102 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.63 | 2.08 | 38 | 5.89 | 2.01 | 0.84 | 0.06 | 0.000216 | |
| 103 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.66 | 2.04 | 38 | 6.63 | 2.57 | 0.44 | 0.06 | 0.001694 | |
| 104 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.63 | 1.78 | 38 | 6.44 | 0.23 | 0.96 | 0.06 | 0.000046 | |
| 105 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.93 | 1.88 | 38 | 5.94 | 2.22 | 0.95 | 0.06 | 0.000056 | |
| 106 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.63 | 2.08 | 33 | 6.36 | 2.21 | 0.58 | 0.06 | 0.000991 | |
| 107 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.66 | 2.04 | 33 | 6.48 | 2.32 | 0.53 | 0.06 | 0.001209 | |
| 108 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.63 | 1.78 | 33 | 6.48 | 2.41 | 0.54 | 0.06 | 0.001197 | |
| 109 | ID41 | T | Indonesia | I | Other | S | W | 33 | 7.93 | 1.88 | 33 | 6.27 | 2.54 | 0.73 | 0.06 | 0.000474 | |
| 110 | ID42 | T | South Korea | O | Other | S | R | 40 | 69.13 | 11.47 | 33 | 66.89 | 11.90 | 0.19 | 0.06 | 0.003217 | |
| 111 | ID42 | T | South Korea | O | Other | S | R | 40 | 70.25 | 12.14 | 33 | 66.89 | 11.90 | 0.28 | 0.06 | 0.002632 | |
| 112 | ID43 | T | China | I | Other | L | S | 29 | 4.16 | 0.46 | 32 | 4.08 | 0.45 | 0.17 | 0.07 | 0.003298 | |
| 113 | ID43 | T | China | I | Other | L | S | 29 | 5.11 | 0.89 | 32 | 4.47 | 0.75 | 0.77 | 0.07 | 0.000372 | |
| 114 | ID43 | T | China | I | Other | L | S | 29 | 3.42 | 0.47 | 32 | 3.49 | 0.43 | −0.15 | 0.07 | 0.005997 | |
| 115 | ID43 | T | China | I | Other | L | S | 29 | 4.59 | 0.76 | 32 | 4.05 | 0.53 | 0.82 | 0.07 | 0.000257 | |
| 116 | ID43 | T | China | I | Other | L | S | 29 | 4.48 | 0.59 | 32 | 4.38 | 0.63 | 0.16 | 0.07 | 0.003385 | |
| 117 | ID43 | T | China | I | Other | L | S | 29 | 4.92 | 0.55 | 32 | 4.57 | 0.59 | 0.60 | 0.07 | 0.000906 | |
| 118 | ID43 | T | China | I | Other | L | S | 29 | 4.92 | 0.49 | 32 | 4.82 | 0.33 | 0.24 | 0.07 | 0.002849 | |
| 119 | ID43 | T | China | I | Other | L | S | 29 | 5.80 | 0.58 | 32 | 5.37 | 0.45 | 0.82 | 0.07 | 0.000252 | |
| 120 | ID44 | S | Egypt | I | ChatGPT | L | V | 31 | 72.11 | 1.66 | 31 | 70.97 | 1.73 | 0.66 | 0.07 | 0.000691 | |
| 121 | ID45 | P | Egypt | B | Other | L | R | 45 | 14.82 | 0.74 | 45 | 8.87 | 1.67 | 4.57 | 0.16 | 0.065844 | † |
| 122 | ID45 | P | Egypt | B | Other | L | R | 45 | 8.00 | 0.85 | 45 | 4.69 | 0.97 | 3.60 | 0.12 | 0.033629 | † |
| 123 | ID45 | P | Egypt | B | Other | L | R | 45 | 8.51 | 0.86 | 45 | 4.49 | 0.94 | 4.42 | 0.15 | 0.060415 | † |
| 124 | ID45 | P | Egypt | B | Other | L | R | 45 | 31.33 | 1.79 | 45 | 18.04 | 2.73 | 5.71 | 0.23 | 0.116510 | † |
| 125 | ID46 | T | Egypt | B | Other | S | W | 30 | 54.37 | 5.54 | 30 | 45.50 | 5.32 | 1.61 | 0.09 | 0.001386 | |

† Studies identified as influential outliers (Cook's d > 0.032).

* Negative constructs (e.g., anxiety, demotivation) were reverse coded so that positive effect sizes here represent improvement in skill.

## Appendix C. Moderator key

Full breakdown of moderators, including those treated collectively due to similarities between them.

| Specific Skill | Terminologies used in the articles |
|---|---|
| Accuracy | "accuracy", "revising", "grammar", "correct words", "omissions"*, "insertions"*, "repetitions"* |
| Reading and Comprehension | "understanding (reading)", "reading comprehension", "comprehension", "reading" |
| Fluency | "fluency", "pronunciation", "accentedness", "comprehensibility" |
| Listening | "listening skills", "listening" |
| Motivation | "interest", "willingness", "motivation", "engagement", "reading engagement" "demotivation"* |
| Organization & Development | "organization", "planning", "cohesion and consistency", "content development", "writing development" |
| Overall speaking | "speaking performance", "speaking", "spontaneous speech", "global speaking" |
| Overall writing | "writing performance", "writing proficiency", "writing" |
| Self-regulation | "self-monitoring", "discipline and regulation", "metacognitive awareness", "cognitive load", "critical thinking", "sense of flow", "self-regulation", "self-corrections", "mindfulness" |
| Self-efficacy | "perceived competence", "confidence", "reduced anxiety", "social emotional competence", "anxiety"*, "listening anxiety"*, "shyness"* |
| Vocabulary knowledge | "lexicon", "productive and receptive vocabulary", "vocabulary knowledge", "lexical quality", "vocabulary" |

*Note: * reverse-coded (effect size multiplied by -1 to accurately present improvement)*

| Mode of Instruction | Description |
|---|---|
| Online | AI integration fully online course |
| Blended | AI used both in- and outside-classroom activities |
| Classroom | AI used in traditional face-to-face instruction |

| **AI Tool** | **Included Tools** |
|---|---|
| ChatGPT | ChatGPT |
| Duolingo | Duolingo |
| Chatbot | "Andy English chatbot", "Google DialogFlow", D-ID Agent, "typebot", "AI-Enhanced WhatsApp", |

(*continued on next page*)

(*continued*)

| AI Tool | Included Tools |
|---|---|
| Other | "Prowritingaid", "Grammarly", "Google Text-to-Speech", "tensor flow", "the English listening learning platform", "Instructor-designed Tool", "Smart Sparrow", "ReadToMe", "ELSA speak", "Microsoft Reading Progress", ReadMate, "Gemini and ChatGPT", "Coze", "Nearpod", "Google Assistant", "Papago and Google AI", "Speechnotes - Speech to Text" |
| **Geographic Area** | **Specific Countries (# of studies)** |
| Other Asia | Afghanistan (2), Kazakhstan (2), Bangladesh (1), India (1) |
| East Asia | China (50), Iran (19), Indonesia (10), South Korea (7), Taiwan (7), Japan (3), Asia Pacific (1), Vietnam (1) |
| Middle East | Oman (8), Egypt (culturally/linguistically) (6), Saudi Arabia (6), Kuwait (1) |
| **Duration** | **Length of study** |
| Long | >10 weeks |
| Short | ≤10 weeks |

## References


Afonso, J., Ramirez-Campillo, R., Clemente, F. M., Büttner, F. C., & Andrade, R. (2024). The perils of misinterpreting and misusing "Publication Bias" in meta-analyses: An education review on funnel plot-based methods. *Sports Medicine, 54*(2), 257–269. https://doi.org/10.1007/s40279-023-01927-9

Ahn, J., & Chi, Y.-K. (2023). 'Are we really back to normal?': A qualitative study on language teachers' teaching practices during the post-COVID-19 era. *Language Teaching Research.* , Article 13621688231205668. https://doi.org/10.1177/13621688231205668

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall, Inc.

Biju, N., Abdelrasheed, N. S. G., Bakiyeva, K., Prasad, K. D. V., & Jember, B. (2024). Which one? AI-assisted language assessment or paper format: An exploration of the impacts on foreign language anxiety, learning attitudes, motivation, and writing performance. *Language Testing in Asia, 14*(1), 45. https://doi.org/10.1186/s40468-024-00322-z

Bognár, L., & Khine, M. S. (2025). The shifting landscape of student engagement: A pre-post semester analysis in AI-enhanced classrooms. *Computers and Education: Artificial Intelligence, 8*, Article 100395. https://doi.org/10.1016/j.caeai.2025.100395

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R., & Higgins, D. J. P. T. (2011). *Introduction to meta-analysis*. Incorporated: John Wiley & Sons.

Canagarajah, S. (2002). *A geopolitics of academic writing*. University of Pittsburgh Press.

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of artificial intelligence in education. *Educational Technology & Society, 25*(1), 28–47.

Chong, S. W., & Plonsky, L. (2021). A primer on qualitative research synthesis in TESOL. *Tesol Quarterly, 55*(3), 1024–1034. https://doi.org/10.1002/tesq.3030

Clark, R. E. (1985). Confounding in educational computing research. *Journal of Educational Computing Research, 1*(2), 137–148. https://doi.org/10.2190/HC3L-G6YD-BAK9-EQB5

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cooper, H. M., Hedges, L. V., Valentine, J. C., & Project Muse (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Harvard Business School Press.

Dong, L., Tang, X., & Wang, X. (2025). Examining the effect of artificial intelligence in relation to students' academic achievement: A meta-analysis. *Computers and Education: Artificial Intelligence, 8*, Article 100400. https://doi.org/10.1016/j.caeai.2025.100400

Dornyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge.

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

Guan, L., Li, S., & Gu, M. M. (2024). AI in informal digital English learning: A meta-analysis of its effectiveness on proficiency, motivation, and self-regulation. *Computers and Education: Artificial Intelligence, 7*, Article 100323. https://doi.org/10.1016/j.caeai.2024.100323

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic Press.

Hsu, T.-C., Chang, C., & Jen, T.-H. (2024). Artificial intelligence image recognition using self-regulation learning strategies: Effects on vocabulary acquisition, learning anxiety, and learning behaviours of English language learners. *Interactive Learning Environments, 32*(6), 3060–3078. https://doi.org/10.1080/10494820.2023.2165508

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society, 26*(1), 112–131.

Hwang, G. J., Rahimi, M., & Fathi, J. (2024). Enhancing EFL learners' speaking skills, foreign language enjoyment, and language-specific grit utilising the affordances of a MALL app: A microgenetic perspective. *Computers & Education, 214*, Article 105015. https://doi.org/10.1016/j.compedu.2024.105015

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 1*, Article 100001. https://doi.org/10.1016/j.caeai.2020.100001

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176*(8), 1091–1096. https://doi.org/10.1503/cmaj.060410

Khan, M. A., Kurbonova, O., Abdullaev, D., Radie, A. H., & Basim, N. (2024). Is AI-assisted assessment liable to evaluate young learners? Parents support, teacher support, immunity, and resilience are in focus in testing vocabulary learning. *Language Testing in Asia, 14*(1), 48. https://doi.org/10.1186/s40468-024-00324-x

Kim, R. (2024). Effects of ChatGPT on Korean EFL learners' main-idea reading comprehension via top-down processing. *Lanaguage Research, 60*(1), 83–106. https://doi.org/10.30961/lr.2024.60.1.83

Kohnke, L. (2024). Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning. *Computers and Education: Artificial Intelligence, 7*, Article 100279. https://doi.org/10.1016/j.caeai.2024.100279

Larsen-Freeman, D. (2015). Saying what we mean: Making a case for 'language acquisition'to become 'language development.'. *Language Teaching, 48*(4), 491–505.

Lee, J.-Y., & Hwang, Y. (2022). A meta-analysis of the effects of using AI chatbot in Korean EFL education. *Studies in English Language and Literature, 48*(1), 213–243. https://doi.org/10.21559/aellk.2022.48.1.011

Li, B., & Peng, M. (2022). Integration of an AI-Based platform and flipped classroom instructional model. *Scientific Programming*, 1–8. https://doi.org/10.1155/2022/2536382, 2022.

Liu, P.-L., & Chen, C.-J. (2023). Using an AI-Based object detection translation application for English vocabulary learning. In *Educational technology & society* (Vol. 26, pp. 5–20). JSTOR. https://doi.org/10.30191/ETS.202307_26(3).0002, 3.

Liu, Z.-M., Hwang, G.-J., Chen, C.-Q., Chen, X.-D., & Ye, X.-D. (2024). Integrating large language models into EFL writing instruction: Effects on performance, self-regulated learning strategies, and motivation. *Computer Assisted Language Learning*, 1–25. https://doi.org/10.1080/09588221.2024.2389923

Mikolajewicz, N., & Komarova, S. V. (2019). Meta-analytic methodology for basic research: A practical guide. *Frontiers in Physiology, 10*, 203. https://doi.org/10.3389/fphys.2019.00203

Moorhouse, B. L., Wong, K. M., & Li, L. (2023). Teaching with technology in the post-pandemic digital age: Technological normalisation and AI-Induced disruptions. *RELC Journal, 54*(2), 311–320. https://doi.org/10.1177/00336882231176929

Piaget, J. (1972). *Psychology and epistemology: Towards a theory of knowledge*. Allen Lane.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

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*, Article 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

Simon, H. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communications, and the public interest*. The John Hopkins Press.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, Article 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ, d4002–d4002*. https://doi.org/10.1136/bmj.d4002

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Torres, P. J. (2021). The role of modals in policies: The US opioid crisis as a case study. *Applied Corpus Linguistics, 1*(3), 1–17. https://doi.org/10.1016/j.acorp.2021.100008

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Viechtbauer, W. (2010). Conducting meta-analyses in *R* with the metafor package. *Journal of Statistical Software, 36*(3). https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology, 55*(1), 10–33. https://doi.org/10.1111/bjet.13334

Zhang, Y., & Dong, C. (2024). Unveiling the dynamic mechanisms of generative AI in English language learning: A hybrid study based on fsQCA and system dynamics. *Behavioral Sciences, 14*(11), 1015. https://doi.org/10.3390/bs14111015

Zhang, S., Shan, C., Lee, J. S. Y., Che, S., & 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

Zimmerman, B. J. (2013). Theories of self-regulated learning and academic achievement: An overview and analysis. *Self-Regulated Learning and Academic Achievement*, 1–36.