From Immediate Performance to Retained Learning: A Multilevel Meta-Analysis of AI-Supported L2 Learning

Mina Song, University of Maryland, College Park, minasong@umd.edu
Shinah Ryu, University of Maryland, College Park, saryu814@umd.edu
Ekaterina Sudina, University of Maryland, College Park, esudina@umd.edu

Although artificial intelligence (AI) has become increasingly prominent in second language (L2) learning, existing syntheses have largely emphasized immediate outcomes and tool-based categories (Wu, 2024; Zhu & Wang, 2025). From a second language acquisition (SLA) perspective, however, immediate performance does not necessarily indicate retained learning, making delayed-posttest evidence essential for evaluating whether AI-supported gains endure over time (Soderstrom & Bjork, 2015). In addition, the same AI tool may serve different instructional roles, such as providing feedback, guiding task completion, supporting interaction, or generating learning materials. Focusing on the pedagogical functions of AI may therefore offer a more theoretically informative lens than tool-based classification alone (Yang & Li, 2024). This meta-analysis examined AI–comparison group differences in delayed-posttest outcomes and tested whether these differences varied by retention interval, AI pedagogical function, and language skill domain.

Eligible studies were identified through systematic searches of ERIC, LLBA, PsycINFO, Scopus, Web of Science, and Google Scholar using search blocks for AI, L2 learning, and delayed-posttest or retention terms. The searches yielded 383 records; after removing 83 duplicates, 300 records were screened, 56 full texts were assessed for eligibility, and 40 studies were included in the final synthesis. Studies were included if they examined an AI-supported L2 intervention with an eligible comparison group, reported a quantitative linguistic performance outcome, included at least one delayed posttest administered at least 5 days after the end of the intervention, and provided sufficient statistical information to calculate Hedges' g effect sizes. The analytic dataset included 163 effect sizes (81 immediate and 82 delayed). Separate effect-size estimates were extracted for each eligible treatment–comparison contrast, outcome, and assessment time point. Multilevel random-effects meta-analytic models were used to account for statistical dependence among multiple effect-size estimates nested within studies. Moderator analyses examined retention interval, AI pedagogical function, and language skill domain, and sensitivity analyses re-estimated the main models after excluding extreme effect-size estimates with $|g| > 4$.

AI-supported groups significantly outperformed comparison groups at the delayed posttest (Hedges' $g = 1.56$, 95% CI [0.89, 2.23], $p < .001$), although substantial heterogeneity was observed. Delayed effects did not differ significantly from immediate effects, suggesting no statistically reliable evidence that AI–comparison group differences changed from immediate to delayed posttest. Moderator analyses did not provide statistically significant evidence that delayed effects varied by retention interval, AI pedagogical function, or language skill domain. Descriptive, category-specific pooled estimates, however, showed positive delayed effects across all coded AI pedagogical functions, including feedback, guided task support, conversational partner, and content generation. Pooled delayed effects were also positive across all examined skill domains, with statistically significant estimates for writing, reading, and vocabulary. In sensitivity analyses, the overall delayed effect remained significant after excluding effect sizes with $|g| > 4$, although the estimate was smaller (Hedges' $g = 1.02$).

Overall, the findings suggest that the advantage of AI-supported groups over comparison groups remained evident in delayed-posttest outcomes, although the current

evidence base was heterogeneous and unevenly distributed. The nonsignificant moderator findings should therefore be interpreted cautiously, given the uneven distribution of studies across functions and skill domains and the frequent co-occurrence of coded categories. By situating AI-supported L2 learning within the SLA distinction between performance and learning, this meta-analysis shifts attention from immediate technological effectiveness to retained L2 development. The study highlights the need for more systematic delayed-posttest designs and more detailed reporting of AI-supported interventions to clarify when, how, and under what conditions AI support is associated with retained L2 learning.

References

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199. https://doi.org/10.1177/1745691615569000

Wu, X.-Y. (2024). Artificial intelligence in L2 learning: A meta-analysis of contextual, instructional, and social-emotional moderators. *System*, *126*, Article 103498. https://doi.org/10.1016/j.system.2024.103498

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System*, *124*, Article 103351. https://doi.org/10.1016/j.system.2024.103351

Zhu, M., & Wang, C. (2025). A systematic review of research on AI in language education: Current status and future implications. *Language Learning & Technology*, *29*(1), 1-29. https://doi.org/10.64152/10125/73606