The Effect of Generative AI (ChatGPT) on EFL Learners' Academic Writing Performance: A Preliminary Meta-Analysis

Budlan Namjil[1]; Khulan Ojgoosh[2]; Tsogzolmaa Guruuchin[3]; Enkhchimeg Tumurkhuyag[4]; Delgermaa Vanchinsuren[5]; Dolzodmaa Namid[6]

[1]Doctoral student, Mongolian National University of Education, Teacher at Mongolian National University

[2]Research manager, University of Engineering and Technology, Mongolia

[3]Department of International Relations, Mongolian National University of Education

[4]Doctoral student, Mongolian National University of Education, Teacher at University of Engineering and Technology, Mongolia

[5]Mongolian University of Life Sciences

[6]Ikh Zasag University, Mongolia

DOI: https://doi.org/10.47772/IJRISS.2026.1026EDU0488

Received: 22 July 2026; Accepted: 27 July 2026; Published: 03 August 2026

ABSTRACT

Generative AI (GenAI), and ChatGPT in particular, is changing how academic writing is taught to learners of English as a foreign language (EFL). Even so, the individual studies and experiments conducted so far report results that conflict with one another, which makes the question worth examining.

Aim. Our study aimed to find out the result of the generative artificial intelligence effect on writing achievement of English language learners quantifying the effect and examine the sources of heterogeneity based on the observed differences.

Methods. To address the research objectives, we implemented a PRISMA 2020-compliant search protocol across Web of Science (n = 99) and ERIC (n = 200). The search incorporated targeted keywords covering topics such as AI writing tools (e.g., ChatGPT, Gemini), metacognitive and personalized learning, EFL academic writing skills, and higher/engineering education. Following a three-phase screening process and full-text assessment, nine eligible empirical studies (N = 624) with between-group post-test comparisons were selected. Standardized effect sizes (Hedges') were extracted and pooled using a restricted maximum likelihood (REML) random-effects meta-analysis.

Nine studies (N=624) were included. The effect was large and positive: g=0.92, 95% CI [0.55, 1.28], p<.001; heterogeneity was high (I^2=79%). Sensitivity analyses were stable (g=0.83 with the AWE tool removed; g=0.88–0.95 across leave-one-out runs). One study (Ekmekçi & Karabulut, 2025) reported a negative effect, where teacher instruction outperformed AI. Secondary constructs (metacognition, learner autonomy) also showed a positive trend. We conclude that GenAI tends to produce a meaningful average improvement in EFL writing performance, but the outcome depends heavily on the conditions of use and on teacher guidance. Because the number of studies is small, the results should be read as preliminary.

Keywords: generative AI, ChatGPT, EFL academic writing, meta-analysis, effect size, PRISMA

INTRODUCTION

Since ChatGPT became publicly available in 2022, research on the use of generative AI (GenAI) in foreign language education, and especially in writing instruction for learners of English as a foreign language (EFL),

has grown sharply. These tools can support the writing process by giving quick feedback, generating ideas, and revising style. Yet individual experimental studies do not add up to a single picture: some report very large positive effects, some report no difference, and a few even report outcomes worse than teacher-led instruction. Under these conflicting conditions, a meta-analysis that systematically combines the quantitative results of individual studies becomes necessary.

This preliminary meta-analysis set out to answer three questions. First, what combined effect does GenAI (ChatGPT and other tools) have on EFL learners' writing performance compared with traditional teaching? Second, how large is the between-study variation in results (heterogeneity), and how far can its causes and contributing factors (tool type, database) explain it? Whatd, beyond writing performance, what effect does GenAI have on underlying learning constructs such as metacognition and learner autonomy?

Because the number of studies is small and the data comes from only two databases, we treat this analysis as a preliminary meta-analysis, and the results are not yet firm enough to be considered reliable.

METHODS

Study design and protocol

We carried out the study as a systematic review and meta-analysis following the PRISMA 2020 reporting guidelines (Page et al., 2021). We note as a limitation that the protocol was not registered in advance.

Inclusion and exclusion criteria (PICOS)

Inclusion criteria: (P) learners of English as a second language; (I) GenAI / ChatGPT or an AI-based writing tool; (C) traditional instruction, teacher feedback, or a control group; (O) writing performance (rubric scores, general writing ability); (S) an experimental or quasi-experimental design with a between-group post-test comparison. **Exclusion criteria:** we excluded studies conducted with native speakers (L1 English), review / conceptual / corpus studies, purely qualitative studies, and studies lacking the quantitative data (M, SD, N or t/F) needed to compute an effect size.

Data sources and search

We searched for the Web of Science (99 records) and ERIC (200 records) in June and July 2026. The search terms covered the AI side (artificial intelligence, ChatGPT, Gemini, AI writing tool), academic writing (academic writing, EFL, foreign language), and the educational context (higher education, university, undergraduate), and we examined international, English-language research articles.

Study selection

We screened titles and abstracts using three relevance levels (directly relevant / relevant / not relevant), then assessed the full text of the empirical articles. Two duplicates were found across the two databases and counted once. The selection flow is shown in Figure 1.

Data extraction and coding

From each study we extracted the author, year, country, tool, design, N of the experimental and control groups, M, SD, the reported statistics (t/F/p), and the effect size. Where a single study reported several sub-outcomes (for example, the three higher-order thinking sub-scores in Hao, Razali, & Zuo, 2026), we assumed a within-study correlation of r=.6 and combined them into a single pooled effect size. When the quantitative data were incomplete, we derived g from the reported d or t.

Effect size and statistical analysis

We computed the between-group standardized mean difference as Hedges' g: $d=(M_1-M_2)/S_{pooled}$; $g=d\times J$, where J is the small-sample correction (Hedges, 1981). A positive g favors AI. We pooled the studies using a random-effects model (REML / Der Simonian–Laird) (Borenstein et al., 2009; Der Simonian & Laird, 1986).

We assessed heterogeneity using I^2, τ^2, and the Q statistic (Higgins & Thompson, 2002). We treated tool type (generative AI [GAI] or automated writing evaluation [AWE]) and database type (Web of Science [WoS] or ERIC) as moderator variables. We also ran a leave-one-out sensitivity analysis and a study-by-study removal sensitivity analysis. All statistics were computed in R using the metaphor package (Viechtbauer, 2010).

Study quality (risk of bias) assessment

We assessed the quality (risk of bias) of the included studies using a version of the Newcastle–Ottawa Scale (NOS) adapted for quasi-experimental designs. The scale has three parts, Selection (up to 4 stars), Comparability (2 stars), and Outcome (3 stars), for a total of 9 stars, and classifies scores as 7–9 = Good, 4–6 = Fair, and 3 or below = Poor. Because the Newcastle–Ottawa Scale was originally designed to assess the quality of observational studies, we adapted it here to fit intervention studies that compare experimental and control groups. The current quality ratings are preliminary and rest on the coded data and article abstracts; indicators such as blinded outcome assessment, independent assessment, and completeness of outcome data (attrition) need to be re-confirmed by two independent raters using the full-text articles. Detailed results are shown in Appendix A, Table A1.

RESULTS

Study Selection Results

From the two databases we found 299 records in total (Web of Science 99, ERIC 200); after removing duplicates (n=2), 297 records were screened by title and abstract. We assessed the full text of 24 empirical candidate records and, after excluding 14, retained 10 studies for the review. Of these, 9 between-group (experimental vs. control) writing-performance studies entered the main meta-analysis, and the remaining study belonged to the secondary (learner autonomy) strand. The selection details are shown in Figure 1.

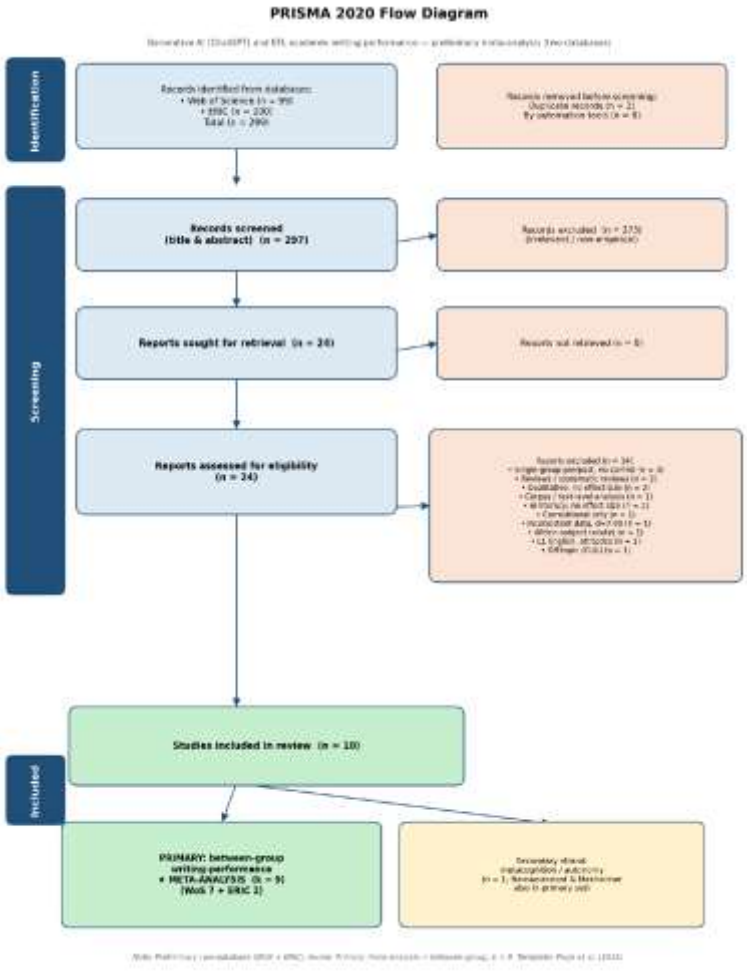

Figure 1. PRISMA 2020 study-selection flow diagram (Web of Science + ERIC).

Selected Articles

The nine included studies covered six countries (Turkey, China, Iran, Indonesia, Vietnam, Egypt) and mostly used quasi-experimental designs with ChatGPT and related GenAI tools. Details are shown in Table 1.

Table 1. Characteristics of the articles included in the meta-analysis (k=9).

Study (year)	Country	Tool	N (E/C)	Outcome	Hedges g [95% CI]
Ekmekçi & Karabulut (2025)	Turkey	ChatGPT etc.	19/21	Writing performance	−0.39 [−1.01, 0.24]
Ozan & Azap (2026)	Turkey	ChatGPT	15/15	Writing performance	0.34 [−0.38, 1.06]
Hao, Razali & Zuo (2026)	China	AWE+GenAI	32/32	Argumentative writing	0.65 [0.22, 1.09]
Hao & Razali (2025)	China	ChatGPT	34/36	Academic writing	0.94 [0.45, 1.44]
Apriani et al. (2025)	Indonesia	ChatGPT	25/25	Academic writing	0.95 [0.36, 1.53]
Namaziandost (2025)	Iran	GenAI	35/35	Writing development	1.11 [0.61, 1.62]
Li & Long (2025)	China	AI assistant	75/75	General ability	1.11 [0.77, 1.46]
Mekheimer (2025)	Egypt	Grammarly (AWE)	30/30	Writing skill	1.61 [1.03, 2.19]
Vo & Ho (2026)	Vietnam	ChatGPT	45/45	Grammatical accuracy	1.69 [1.21, 2.18]

Pooled Effect

When pooled with the random-effects model, GenAI showed a large, positive effect on EFL learners' writing performance: g=0.92, 95% CI [0.55, 1.28], p<.001 (k=9). The individual and pooled effect sizes are shown in Figure 2.

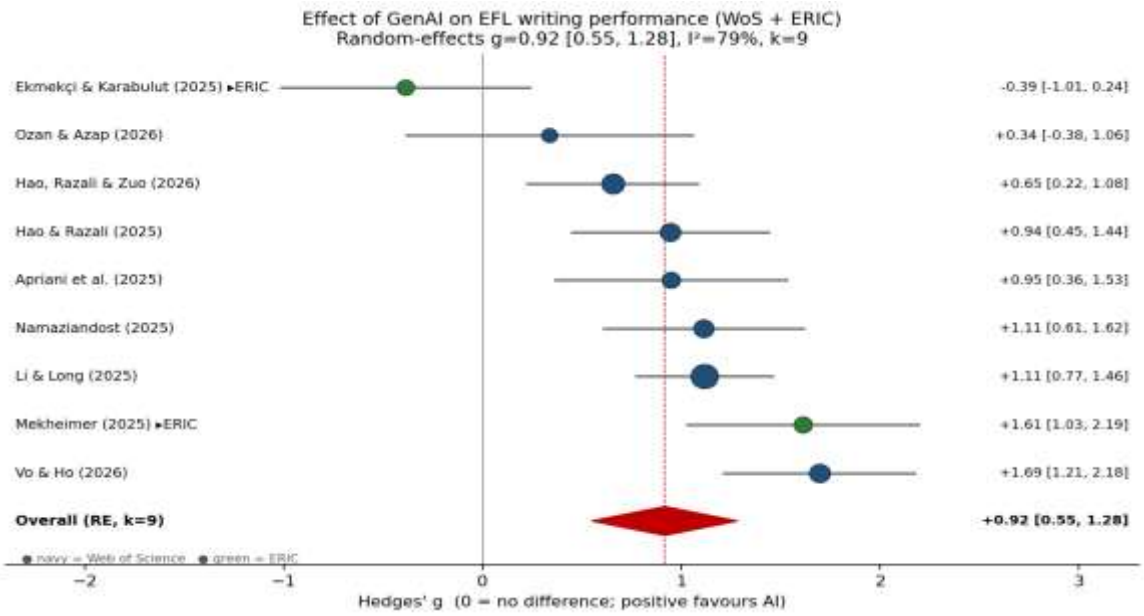

Figure 2. Forest plot of between-group Hedges' g (k=9). Dark blue = Web of Science, green = ERIC.

Heterogeneity and Moderator Analysis

Heterogeneity was high (I^2=79%, τ^2=0.24, Q (8) =37.4, p<.001), meaning the true variation between studies is large. In the subgroup analysis, the Web of Science studies (k=7) gave g=1.00 [0.72, 1.29] (I^2=58%), while the ERIC studies (k=2) gave g=0.62 [−1.34, 2.58] (I^2=95%); the latter consists of only two studies with opposite results, so it must be interpreted with caution. The tool-type moderator (GAI vs. AWE) is preliminary because the number of studies is small.

Sensitivity Analysis

The results were relatively stable. Removing Mekheimer (2025), which used an AWE tool (Grammarly), gave g=0.83 [0.45, 1.22]; removing Li & Long (2025), noted as possibly having an inflated effect, gave g=0.88 [0.45, 1.32]; using only the pure GAI studies gave g=0.83. In the leave-one-out analysis, the pooled g did not change substantially in any single case. Ekmekçi & Karabulut (2025) was the only study with a negative effect and the main source of heterogeneity.

Publication Bias

Because the number of studies is fewer than 10 (k=9), funnel-plot symmetry and Egger's test are not reliable; Figure 3 is shown only as a visual reference. There are too few studies to draw a formal conclusion about bias.

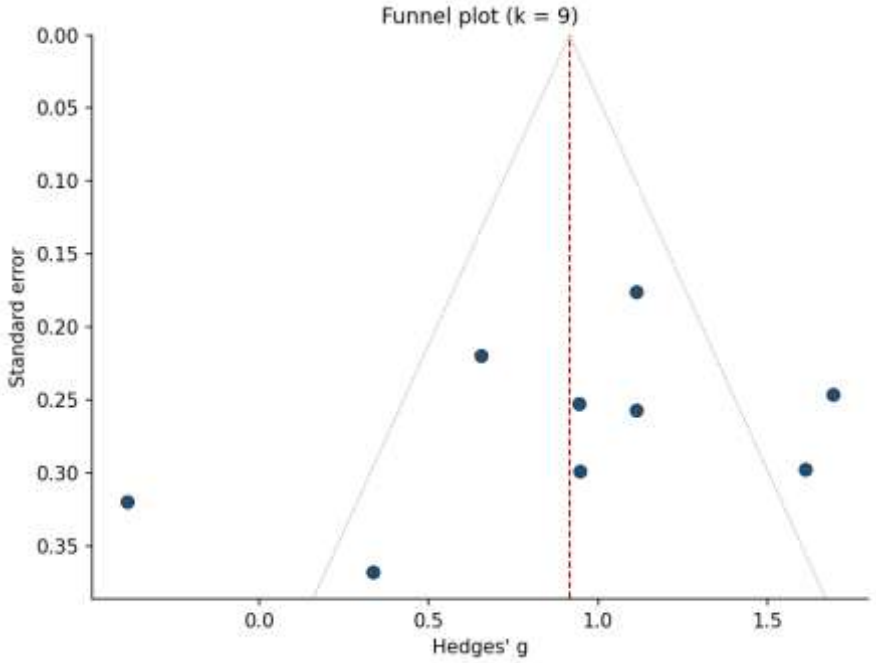

Figure 3. Funnel plot, k=9. The dashed red line marks the pooled effect size.

Secondary and Correlational Strands

Beyond writing performance, a positive trend also appeared in other constructs. For metacognitive awareness, Namaziandost (2025) reported a large effect (g≈1.05). For learner autonomy, Nguyen & Doan (2025) found a moderate positive effect (g=0.82), but no difference in self-efficacy (g=0.28, not statistically significant). In the correlational strand, Mekheimer (2025) found a positive relationship between the frequency of using GenAI features and writing scores (r≈.35–.42). Because these strands include few studies, they are described without pooling.

Study Quality

By the preliminary NOS assessment, the studies scored 4–7 out of 9 (mean ≈ 6.0); four were rated "Good," five "Fair," and none "Poor." The most common uncertainties were blinded / independent outcome assessment (O1) and completeness of outcome data / attrition (O3). Ekmekçi & Karabulut (2025) showed a discrepancy

in sample reporting (O3), and Li & Long (2025) showed a risk in the outcome–intervention match. Overall, the quality was fair to good, which relatively supports the pooled result, but given the high heterogeneity and the small number of studies, it should be interpreted with caution. Details are shown in Appendix A.

DISCUSSION

The main finding of this preliminary meta-analysis is that GenAI (ChatGPT and related tools) tends to improve EFL learners' writing performance by a large average amount ($g \approx 0.92$). This agrees with the conclusions of qualitative and systematic reviews, but the high heterogeneity ($I^2=79\%$) is a reminder that the result cannot be treated as a fixed effect estimate and that its causes need to be accounted for.

A case that stands out is the negative result of Ekmekçi & Karabulut (2025): in their study, teacher instruction was more effective than GenAI, and the experimental group made no gains. This shows that the effect of GenAI depends less on the tool itself than on how the learner manages and uses it and what guidance they receive. Used without guidance, learners risk over-relying on it and handing self-monitoring over to the tool, a "metacognitive laziness" (Fan et al., 2025). Metacognitive support and appropriate teacher involvement therefore play a decisive role.

As for tool type, even when the AWE tool (Grammarly), which gives only corrective feedback, is removed, the pooled effect stays large ($g=0.83$), which indicates that the result does not depend on a single type of tool. Pedagogically, stabilizing positive learning impacts requires positioning GenAI as an interactive learning instrument-backed by guided instruction—rather than a standalone query responder, thereby enhancing metacognitive awareness and feedback literacy.

Finally, existing studies largely overlook the influence of L1 linguistic backgrounds, cultural context, and contrastive rhetoric (Kaplan, 1966). This unexamined domain presents a significant research opportunity to explore how GenAI interacts with rhetorical development in academic writing.

LIMITATIONS

This analysis has several limitations. First, because the number of studies is small ($k=9$), the pooled figure is fragile, and publication bias cannot be checked reliably. Second, because it covers only two databases and English-language research articles, selection bias may arise. Third, the protocol was not registered in advance. Fourth, heterogeneity is high ($I^2=79\%$). Fifth, some effect sizes were derived from reported d/t, and the post-test SMD does not fully capture the differences of ANCOVA-adjusted models. Sixth, the effect size of one study (Li & Long, 2025) may be inflated by test–task alignment, and another (Ekmekçi & Karabulut, 2025) had a noted sample discrepancy. For these reasons, it is better to view this as a preliminary analysis that identifies a trend rather than one that produces a definitive pooled result.

CONCLUSION

This preliminary meta-analysis shows that GenAI (ChatGPT) tends to produce a meaningful average improvement in EFL learners' academic writing performance ($g \approx 0.92$). The effect, however, varies considerably between studies and depends on the conditions of use, teacher guidance, and learner metacognition. Going forward, it will be necessary to add longer-term studies from other databases (such as Scopus), increase the number of studies to $k \geq 12$, and analyze the sources of heterogeneity more deeply through moderators.

REFERENCES

1. Apriani, E., Daulay, S. H., Aprilia, F., Marzuki, A. G., Warsah, I., Supardan, D., & Muthmainnah. (2025). A mixed-method study on the effectiveness of using ChatGPT in academic writing and students' perceived experiences. Journal of Language and Education, 11(1), 26–45. https://doi.org/10.17323/jle.2025.17913

2. Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Wiley.
3. DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2
4. Ekmekçi, E., & Karabulut, E. (2025). Chatbot-driven writing practice: A boon or a bust for EFL learners? GIST Education and Learning Research Journal, (31), 45–64.
5. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544
6. Hao, H., & Razali, A. B. (2025). Impact of ChatGPT on English academic writing ability and engagement of Chinese EFL undergraduates. International Journal of Instruction, 18(2), 323–346. https://doi.org/10.29333/iji.2025.18218a
7. Hao, H., Razali, A. B., & Zuo, R. (2026). Exploring the impact of integrated AWE and generative AI feedback on Chinese EFL undergraduates' higher-order thinking in argumentative writing. SAGE Open, 16(1). https://doi.org/10.1177/21582440251413884
8. Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107–128. https://doi.org/10.3102/10769986006002107
9. Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539–1558. https://doi.org/10.1002/sim.1186
10. Li, C., & Long, J. (2025). Comparing AI-assisted and traditional teaching in college English: Pedagogical benefits and learning behaviors. Information, 16(10), Article 895. https://doi.org/10.3390/info16100895
11. Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. Discover Education, 4, Article 170. https://doi.org/10.1007/s44217-025-00602-7
12. Namaziandost, E. (2025). Integrating flipped learning in AI-enhanced language learning: Mapping the effects on metacognitive awareness, writing development, and foreign language learning boredom. Computers and Education: Artificial Intelligence, 9, Article 100446. https://doi.org/10.1016/j.caeai.2025.100446
13. Nguyen, Q. N., & Doan, D. T. H. (2025). ChatGPT's effects on EFL learners' autonomy and self-efficacy in self-regulated learning contexts. The JALT CALL Journal, 21(2), Article 102968.
14. Ozan, O., & Azap, S. (2026). Employing ChatGPT to improve high school students' writing skills by providing feedback on topic-specific writing tasks. Computer Assisted Language Learning. Advance online publication. https://doi.org/10.1080/09588221.2026.2613219
15. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
16. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03
17. Vo, T. L., & Ho, T. M. P. (2026). A quasi-experimental study on enhancing EFL learners' grammar precision in writing skills under supported-ChatGPT approach at a university in a remote region of Vietnam. Arab World English Journal, 361–375. https://doi.org/10.24093/awej/A3.23

APPENDIX A. STUDY QUALITY ASSESSMENT (ADAPTED NEWCASTLE–OTTAWA SCALE)

Table A1. Preliminary quality (risk of bias) assessment of the 9 included studies.

Study (year)	Selection /4	Comp. /2	Outcome /3	Total /9	Rating	Note
Ekmekçi & Karabulut (2025)	4	1	1	6	Fair	Sample reporting discrepancy (O3)
Ozan & Azap (2026)	4	1	1	6	Fair	O1/O3 unclear
Hao, Razali & Zuo (2026)	4	2	1	7	Good	Baseline controlled via ANCOVA
Hao & Razali (2025)	3	0	1	4	Fair	Baseline balance unclear
Apriani et al. (2025)	4	0	1	5	Fair	Baseline balance reporting unclear
Namaziandost (2025)	4	2	1	7	Good	Baseline controlled via ANCOVA
Li & Long (2025)	4	1	2	7	Good	Outcome–intervention match risk
Mekheimer (2025)	4	1	2	7	Good	Objective proficiency test
Vo & Ho (2026)	3	0	2	5	Fair	Baseline balance unclear

Note. The NOS was adapted for quasi-experimental designs. Selection (max 4), Comparability (max 2), Outcome (max 3). 7–9 = Good, 4–6 = Fair, ≤3 = Poor. Ratings are preliminary and based on coding and abstracts; blinded / independent assessment (O1) and attrition (O3) should be confirmed from the full texts by two raters.