Research review of Zhang et al. (2026)

When does GenAI work for L2 learning? A three-level meta-analysis

Review of Education | DOI: 10.1002/rev3.70252 | 14 November 2026

SUMMARY

Zhang et al. combined 48 studies on generative AI in second-language learning, yielding 67 effect sizes. Their pooled estimate is g = 0.76; it falls to g = 0.57 when three unusually large effects are removed. The results vary considerably across studies. A separate check of the published tables also reveals an unexplained difference of 818 participants: Table 2 sums to 4,378, while the main text reports 5,196. Questions remain about the inclusion of several studies, the classification of one study design and two statistical calculations. The evidence points to better performance in some instructional settings. The size and causes of that benefit, and what students can later do independently, are less certain.

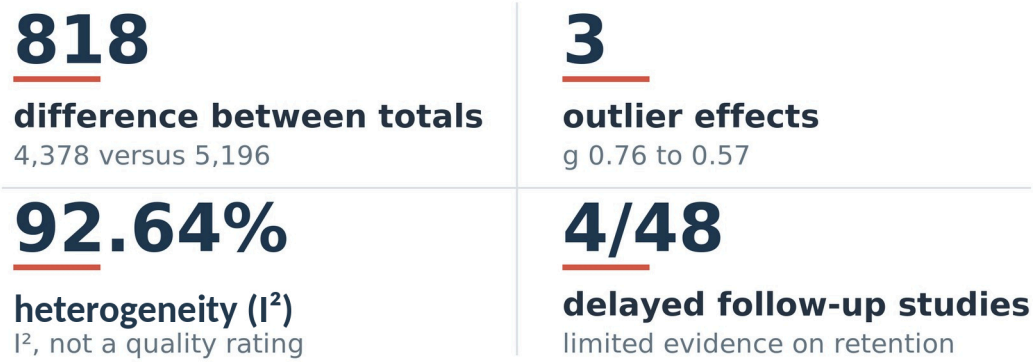


Figure 1. Four findings that matter when interpreting Zhang et al. [1]

What did Zhang and colleagues investigate?

Zhang, Aubrey, Wu and Chiu examine whether generative AI can help students learn a second or foreign language. They also ask whether outcomes differ according to how students and teachers use it. Their meta-analysis draws together studies of writing, speaking, reading and vocabulary. AI plays different roles across these studies, from conversational partner to feedback provider. These distinctions matter: practising a conversation with a chatbot is not the same learning activity as revising a text using automated feedback. [1, pp. 1–6]

The authors searched six scholarly databases for work published between 2023 and February 2026. They assessed 458 full-text articles and included 48 studies, which yielded 67 effect sizes. An effect size places differences between groups on a common scale, allowing results from different assessments to be compared. Some studies contributed multiple outcomes. A three-level model accounts for the dependence between results from the same study. This is a sensible choice: two tests from one classroom do not represent two independent experiments. [1, pp. 7–16]

The paper does more than present an average. It compares language skills, AI applications and teaching arrangements. The authors also test how the estimate changes when extreme results are excluded and look for possible publication bias. Such checks are appropriate in a body of research this varied. [1, pp. 14–20]

What do the results show?

The authors report Hedges' g = 0.76, with a 95% confidence interval of 0.50 to 1.02. Across the measured language outcomes, AI-supported groups generally scored higher than comparison groups. The figure does not mean a 76% improvement. It is a standardized difference between groups and does not tell us how much any individual student learned. [1, p. 17]

The estimate varies by skill. The authors report g = 1.48 for speaking, 0.66 for vocabulary and 0.61 for writing. For reading, the estimate is g = 0.076 and is not statistically significant. The five reading studies therefore do not provide convincing evidence of a positive pooled effect. That does not establish that AI cannot help with reading instruction. [1, p. 18]

Three exceptionally large effects exert considerable influence: 5.70 and 4.56 from Elmaadaway et al. (2025), and 5.16 from Wang (2025). Excluding them reduces the overall estimate from 0.76 to 0.57. The estimate for speaking drops from 1.48 to 0.65 and that for AI feedback from 1.15 to 0.83. The result remains positive in this sensitivity analysis. Its magnitude nevertheless changes substantially when only three effects are removed. [1, p. 20]

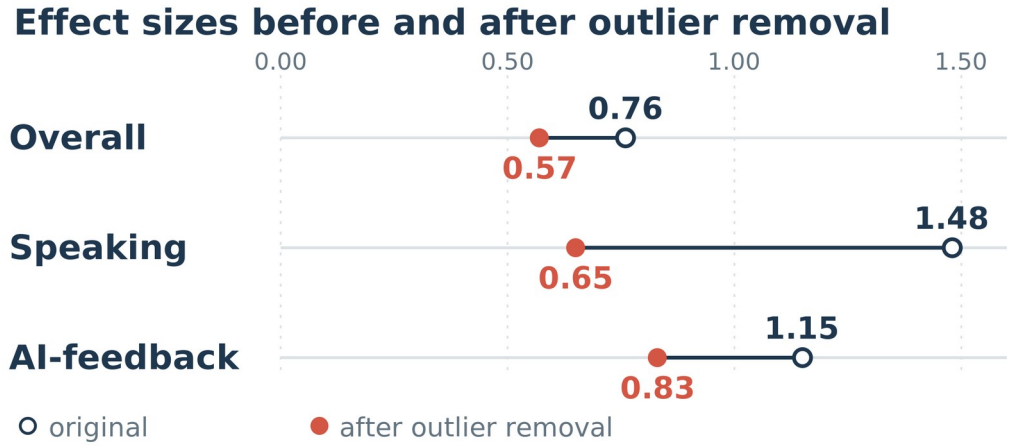


Figure 2. Published estimates on a common scale. Open circles: full analysis. Red dots: after outlier exclusion. [1, p. 20]

Why do the effects differ so much?

The authors report I^2 = 92.64%. The measured effects vary far more than sampling error alone would lead us to expect. This is not a claim that 92.64% of the studies are poor quality. It explains why one pooled estimate tells us relatively little about what would happen in a particular classroom. [1, p. 17]

There are also signs of small-study effects. The cluster-robust Egger test gives p = 0.006. One possible explanation is publication bias, for example if positive studies are more likely to appear in journals. Other explanations remain possible. The test does not identify any specific unpublished research. [1, pp. 17, 23]

Do all included studies fit the research question?

A meta-analysis depends on which studies are included. Zhang et al. require research on second-language learning, a measured language outcome and a comparison between groups. The intervention must also generate language rather than simply

retrieve pre-written responses. Yet the fit of several included studies with those criteria remains unclear. [1, pp. 7–8]

Bašić et al. (2023) asked eighteen forensic-science master's students to write an essay about biometric identification. They wrote in Croatian, their first language. The authors measured essay quality, writing time and authenticity with and without ChatGPT. This provides evidence about writing assistance. The article does not establish that the students were learning a second language. Its inclusion therefore needs an explanation. [1, p. 9; 2]

Avello et al. (2024) examined whether ChatGPT helped master's students understand digital storytelling and reduced cognitive load. Zhang et al. count this study under writing. Yet knowledge about digital storytelling is not, by itself, an outcome measure of second-language acquisition. Which language skill was measured remains unclear. [1, pp. 8–9; 3]

Aydın Yıldız (2023) presents a different case. Motivation and learning strategies feature prominently, but the study also includes a vocabulary test. It cannot simply be ruled out. The question is which result Zhang et al. used: the vocabulary score or a questionnaire that did not measure language ability. [1, p. 9; 4]

The technology also needs checking. Elmaadaway et al. (2025) studied reading aloud with Amazon Alexa. Hwang et al. (2023) used Smart RoamLingo, which combines recognition technologies with feedback. The label "AI" alone does not establish that a language model generated the relevant responses. That must be demonstrated for each intervention. The Alexa question is especially consequential because two of the three extreme effects come from that study. [1, pp. 7–10, 20; 5, 6]

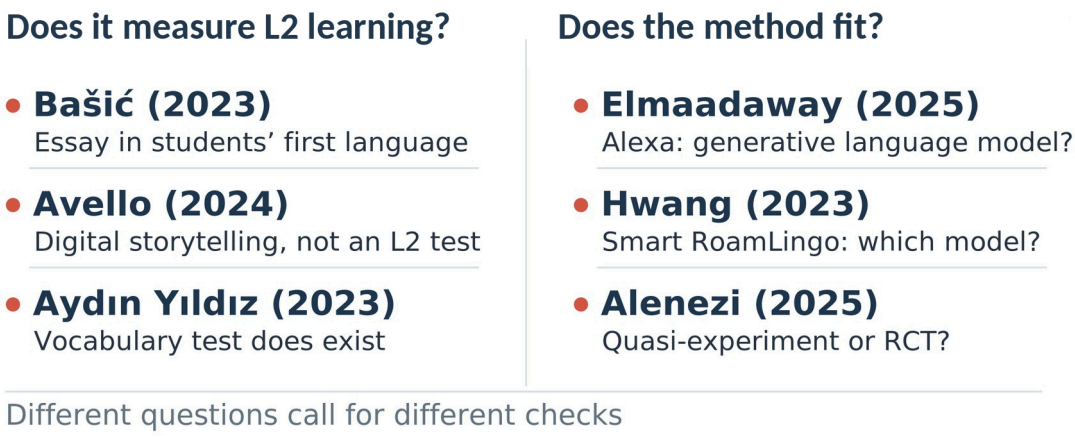


Figure 3. Six studies require different checks of their eligibility.
Those concern the learning outcome, technology or study design. [1–7]

How strong are the comparisons?

Table 2 classifies 25 studies as randomized experiments and 23 as quasi-experimental. Alenezi and Alenezi (2025) is included in the randomized group. The primary authors, however, describe their work as quasi-experimental and report purposive selection of thirty secondary-school students. The article inspected does not show that participants were randomly allocated to the groups. [1, pp. 9, 16; 7]

This matters beyond terminology. Zhang et al. assess risk of bias with RoB 2 for randomized trials and ROBINS-I for non-randomized studies. Two reviewers conducted

the assessments independently and, according to the authors, agreed on 85% of their decisions. For Alenezi, it remains to be established which instrument was actually applied and whether it matched the study design. [1, p. 15; 7]

The classification also enters the results. The pooled estimate is g = 1.02 for randomized studies and g = 0.48 for quasi-experimental ones. Misclassifying a study may change that comparison. By how much can only be determined after checking the allocation procedures and rerunning the analysis. [1, p. 19]

The comparison conditions deserve equal attention. In Meyer et al. (2024), 459 students completed an English writing task. The AI group received language-model feedback and could revise their texts accordingly; the comparison group revised without that feedback. The AI group had a small advantage (d = 0.19). That finding is meaningful for this specific contrast. It does not establish what AI adds compared with an equal amount of feedback from a teacher. [8]

In other studies, introducing AI also changed the amount of feedback, practice time or instruction. Any measured advantage may therefore reflect the entire instructional arrangement, rather than the contribution of the language model alone. Zhang et al. acknowledge this limitation. [1, p. 23; 9]

Why do the participant counts not agree?

The results section reports 5,196 participants, comprising 2,584 in the experimental groups and 2,612 in the control groups. Table 2 lists a sample size for each of the 48 studies. Those entries sum to 4,378. The table spans pages 9–12, with Table 3 beginning on page 13. The page totals are 1,041, 1,689, 1,261 and 387, leaving a difference of 818 from the reported total. [1, pp. 9–12, 16]

The paper does not explain the difference of 818. The summary table might contain an error, or the two totals might follow different counting rules. Sample sizes affect how much weight a study receives in the model. The group sizes actually used are needed to determine whether this is a reporting discrepancy or also affects the analysis. [1]

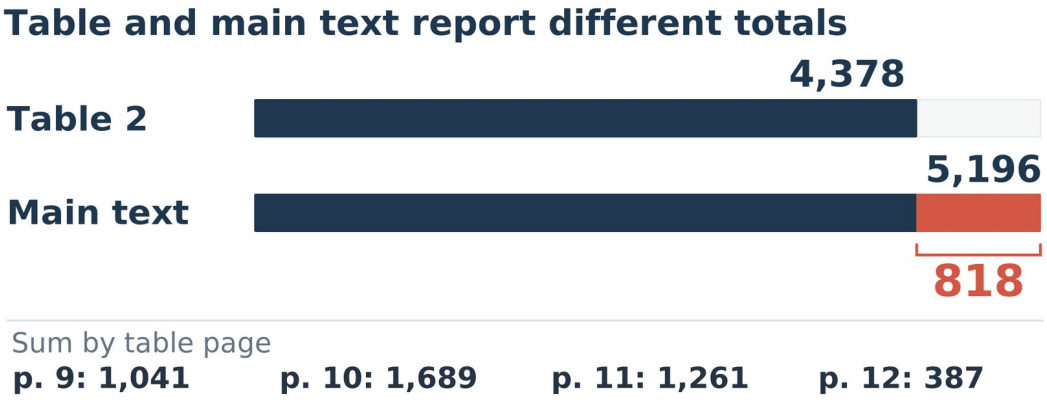


Figure 4. Red marks only the difference between two published totals. It does not represent an additional group of participants. [1]

What do the recalculations show?

One of the largest effects comes from the individual study by Y. Wang (2025). It reports groups of 33 and 35 students, means of 3.66 and 1.27, and standard deviations of 0.38 and 0.52. These numbers yield Hedges' g ≈ 5.16, as Zhang reports. But a conventional independent-groups t test using the same figures gives approximately

21.5, whereas the source reports t = 9.72. The reported value may relate to a different test or outcome. Without clarification, the two figures cannot be reconstructed as the same comparison. [1, p. 20; 10]

The moderator analysis in Zhang's own article also contains a mismatch. It reports Q(4) = 10.37 and p = 0.016. With four degrees of freedom, that Q-value corresponds to approximately p = 0.035. Three degrees of freedom give about 0.016 and would be usual for four categories. The discrepancy may therefore be a notation error. The original model output is needed to establish which interpretation is correct. [1, p. 19]

Neither mismatch proves that the overall results are wrong. We need the actual test definitions, outcome variables and model settings to determine what happened. The published figures alone do not provide all of that information.

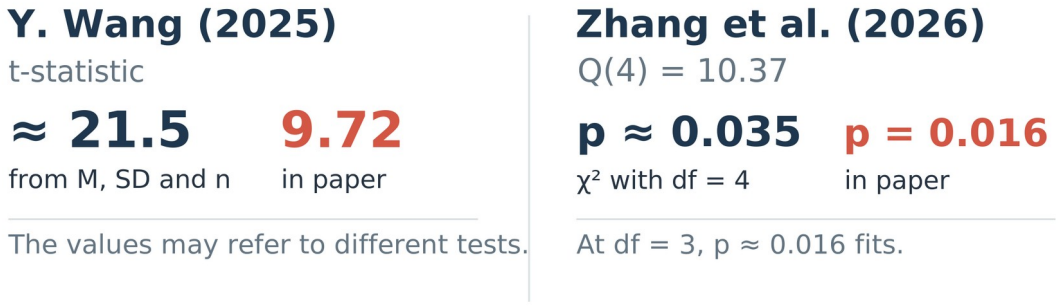


Figure 5. Calculated values compared with those reported in the original publications. [1, 10]

To whom do the findings apply?

The 48 studies do not represent all language-learning settings. Thirty-nine took place in higher education; twenty-six were conducted in China, including Hong Kong, Macau and Taiwan. Only thirteen used standardized language assessments, according to Zhang. Other studies employed tests developed by their researchers. Such tests are not necessarily poor, but differences between measures complicate comparisons across studies. [1, p. 16]

For their proficiency categories, the authors placed, among others, Egyptian fourth graders and Chinese sixth graders in the beginner group (CEFR A1–A2 or equivalent). They based this decision on developmental stage and instructional context, without reporting a measured CEFR level for those particular groups. The proficiency comparisons therefore do not rely on the same kind of evidence throughout. [1, pp. 8, 13]

The evidence on intervention duration is thin. Data on instructional hours came from only 17 studies, yielding 24 effect sizes. The authors found no statistically convincing relationship between duration and outcome. That is not evidence that duration makes no difference. The analysis rests on a limited subset of the studies. [1, p. 19]

Only four of the 48 studies included a delayed post-test. Most outcomes were measured during or soon after instruction. Such tests can certainly provide evidence of learning, especially if students demonstrate the skill independently. But a text improved with AI feedback during practice does not show that the same student can write equally well without help weeks later. Applying the skill to a different task also requires separate evidence. [1, pp. 16, 23; 8]

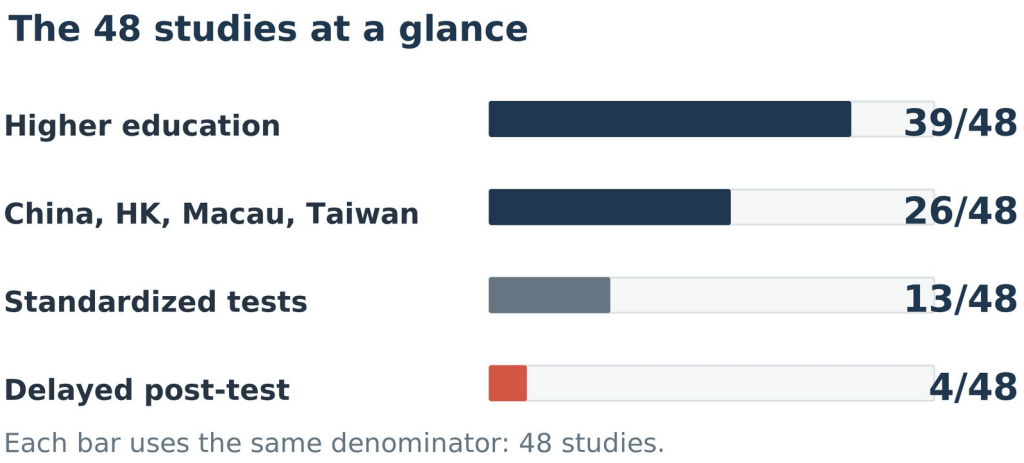


Figure 6. Four characteristics of the 48 studies. The categories overlap; they are not mutually exclusive. [1, p. 16]

Zhang et al. compare how initiative is divided between the learner and AI. They report g = 0.92 for learner-led use, 0.82 for co-creative use, 0.58 for instrumental use and 0.34 for AI-led use. These estimates, however, come from different sets of studies. There is no single experiment in which only the distribution of initiative was changed. The association therefore does not establish that greater learner control caused the stronger outcomes. [1, p. 19]

The authors assigned these categories mainly from published accounts of the lessons. Two coders agreed closely (Cohen's κ = 0.86), indicating consistent coding. That does not show who initiated actions, noticed mistakes, weighed feedback or decided how to rephrase an answer. Those actions were not directly observed. [1, pp. 8, 14, 21]

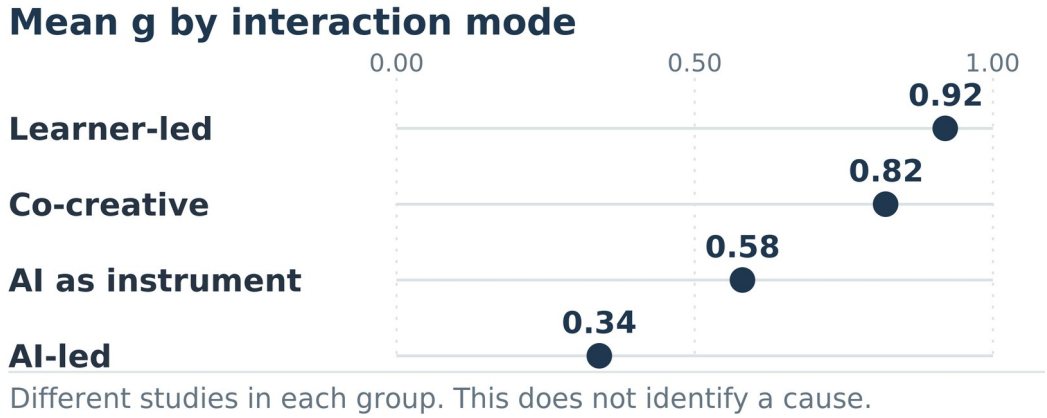


Figure 7. Pooled effects by interaction mode. The categories contain different studies, not an experiment that varied only task ownership. [1, p. 19]

The main analysis did not find a convincing difference between teacher-role categories. A more observational teacher role appeared more favorable only after the outliers were excluded. The authors describe the result as exploratory. It cannot support a general recommendation for teachers to step back. [1, pp. 19–23]

What can we conclude from this study?

The average remains positive even after the three extreme effects are excluded. That result deserves recognition. The three-level model suits the data structure, and the authors are transparent about the variation between studies. Their checks for outliers

and possible publication bias make the findings easier to assess. In several of the settings examined, AI-supported students scored higher on the measured language outcomes. [1]

The precise size of that advantage cannot be independently reproduced from the published tables. Doing so requires all 67 effect sizes, along with group sizes, variability measures, study designs and the AI applications used. We could then test the consequences of changing the source selection or correcting Alenezi's classification. Those consequences cannot yet be quantified.

A second question concerns what those assessments measured. A higher score is a meaningful finding. To know what a student can do independently, we need to know how the assessment was conducted and whether AI support remained available. Evidence on retention over time and application to new tasks is scarce in this meta-analysis. A general claim about lasting language proficiency goes beyond what the studies establish.

What is needed to verify the findings?

Zhang et al. state that all analyzed data are included in the article and that additional material is available on request. However, the 67 effect sizes and standard errors do not appear alongside all group sizes, risk-of-bias assessments and model variables in a fully reproducible dataset. The authors name the software but provide no runnable analysis code with the article examined here. The tables alone are insufficient for independent replication. [1, pp. 15, 24]

With those materials, the analysis could be rerun using the original study selection and then repeated with disputed studies excluded individually and study designs reassessed. That would show whether the discrepancies concern the published account alone or also change the pooled result. For now, the positive average is a reported finding whose precise magnitude remains unverified.

- All 67 effect sizes, including standard errors, within-study dependence and the group sizes actually used.
- A complete list of included and excluded studies, with the reason for each decision and the selected language outcome.
- Technical documentation of the Alexa application and Smart RoamLingo showing whether they meet the GenAI eligibility criteria.
- Study-design and interaction-mode coding, including the justification for classifying Alenezi as randomized.
- The analysis code and results of reruns, both without the three outliers and with separate sensitivity analyses for disputed studies.

References

[1] Zhang, Z., Aubrey, S., Wu, X., & Chiu, T. K. F. (2026). When does GenAI work for L2 learning? A three-level meta-analysis. Review of Education, 14, e70252. DOI

[2] Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. Humanities & Social Sciences Communications. DOI

[3] Avello, R., Gajderowicz, T., & Gómez-Rodríguez, V. G. (2024). Is ChatGPT helpful for graduate students in acquiring knowledge about digital storytelling and reducing their cognitive load? Revista de Educación a Distancia. DOI

[4] Aydın Yıldız, T. (2023). The impact of ChatGPT on language learners' motivation. Journal of Teacher Education and Lifelong Learning, 5(2), 582–597. DOI

[5] Elmaadaway, M. A. N., et al. (2025). Improving primary school students' oral reading fluency through voice chatbot-based AI. Journal of Computer Assisted Learning. DOI

[6] Hwang, W.-Y., et al. (2023). AI and recognition technologies to facilitate English as foreign language writing. Journal of Educational Computing Research. DOI

[7] Alenezi, A., & Alenezi, A. (2025). Evaluating the effectiveness of chatbot-assisted learning in enhancing English conversational skills. Education Sciences, 15(9), 1136. DOI

[8] Meyer, J., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom. Computers and Education: Artificial Intelligence, 6, 100199. DOI

[9] Lawson, A. P., et al. (2026). Color me confounded: A critical analysis of media comparisons on ChatGPT in education. Computers & Education, 254, 105701. DOI

[10] Wang, Y. (2025). A study on the efficacy of ChatGPT-4 in enhancing students' English communication skills. SAGE Open, 15(1). DOI

Calculation notes and outstanding checks

Participants: p. 9 = 1,041 (14 studies); p. 10 = 1,689 (13); p. 11 = 1,261 (15); p. 12 = 387 (6). 48 studies, 4,378 participants. The main text reports N = 5,196; difference 818. Table 3 begins on p. 13. [1]

Wang (2025): $n_1 = 33$, $n_2 = 35$, $M_1 = 3.66$, $M_2 = 1.27$, $SD_1 = 0.38$, $SD_2 = 0.52$. Pooled SD ≈ 0.4575; d ≈ 5.224; Hedges' g ≈ 5.164; conventional independent-group t ≈ 21.5. The source also reports t = 9.72; clarification is needed on which test was performed. [10]

Interaction moderator in Zhang (p. 19): Q(4) = 10.37, p = 0.016 as printed. A χ^2 test with df = 4 yields p ≈ 0.0346, whereas df = 3 yields p ≈ 0.0157. [1]

This is a targeted audit of published statistics and selected source papers, not a full reconstruction of all 67 effect sizes. EAI Core 0.5.4 is used only to distinguish performance, independence, retention and transfer.

EAI disclosure: how this review was produced

Hans Visser and ChatGPT collaborated on this research review. Its analysis, writing and presentation were developed jointly. This account reflects Visser's description, not a complete record of the tools used.

Task	Hans Visser	ChatGPT
Research	• select sources	• search and organise
Analysis	• calculate, test, interpret	• recalculate, compare
Text and figures	• rewrote much text	• drafts and diagrams

Final review and editorial responsibility: Hans Visser

Figure 8. Analysis and writing were shared activities. Hans Visser carried out the final substantive review.

Source research. Hans Visser defined the questions, selected sources and assessed their relevance. According to the workflow he described, the Exa agent supported searches for original publications, while Firecrawl was used to extract material from public web pages. ChatGPT helped organise and compare the retrieved sources.

Analysis. Hans Visser and ChatGPT examined the methodology, source selection, calculations and conclusions together. Visser contributed his own analyses and checked the interpretation. ChatGPT helped recalculate published statistics and compare source material. Substantive judgments were developed collaboratively.

Writing and editing. The structure and prose were developed jointly. Hans Visser extensively rewrote the text and guided the reasoning, style and tone. ChatGPT contributed drafts, editorial suggestions, dual-coding diagrams and document design. Visser conducted the final substantive review and retained editorial responsibility.

Disclosure. AI tools are not independent evidence for research claims; those claims must be checked against the cited publications and calculations. The workflow is described on the author's account, not from a complete tool log. All 67 effect sizes in Zhang et al. have not been independently recalculated.