

Presumed Effective: The Manufacturing of an Evidence Base for AI-in-Education Through Flawed

Meta-Analysis

Dr. Patrick E. O'Neill

Ivy Tech Community College

poneill6@ivytech.edu

ORCID: <https://orcid.org/0009-0003-7795-5817>

Preprint Version - This manuscript has not yet undergone peer review.

Abstract

This review employs a forensic methodological audit of 14 meta-analyses making broad claims about the impact of AI on education, evaluating construct coherence, primary study selection, and statistical validity. A brief review of meta-analytical methods establishes the best practice benchmarks used in the audit. The audit found that none of the examined meta-analyses provided a valid basis for the claims they advanced: none had a coherent construct, none sufficiently assessed publication bias, and all had severe heterogeneity. Statistics were misapplied, miscalculated, and misinterpreted. A majority (61%) of the randomly vetted primary studies were problematic, most commonly because the outcome measured didn't match the meta-analysis. Publication bias is well-documented in education literature, especially in edtech. Given a positively selected education literature, a meta-analysis with weak screening and inadequate publication-bias correction is aggregating publication bias and calling it evidence. These failures are consequential in a policy environment where AI adoption is increasingly treated as urgent and inevitable. The audited meta-analyses are heavily cited, with over 2,000 combined citations despite having been published for only 16 months on average as of this writing. These findings show how AIED meta-analysis can manufacture certainty from an evidence base that is heterogeneous, positively selected, and weakly vetted. The recurrence of severe validity failures across peer-reviewed journals points to a failure of editorial and reviewer gatekeeping, not isolated author error. Requirements for data sharing and stronger editorial and reviewer standards are recommended for meta-analyses in education journals.

Keywords: artificial intelligence in education; meta-analysis; publication bias; education technology; research synthesis; academic integrity; educational policy

Presumed Effective: The Manufacturing of an Evidence Base for AI-in-Education

1. Introduction

1.1 The rapid adoption of AI in education

Artificial intelligence in education (AIED) has moved faster than researchers can provide evidence to guide it (Reich, 2025). Since the public release of ChatGPT, educational institutions, policymakers, and technology firms have treated AI not as a hypothetical future intervention but as an immediate necessity. There is a longstanding pattern of optimism toward new technology in education, with recommendations for adoption often preceding evidence of benefit (Higgins et al., 2012), but this push is structurally and economically different from the edtech hype cycles of the past. We are currently witnessing one of the largest deployments of private capital in modern economic history with US cloud and AI infrastructure providers on track to spend \$700 billion in 2026 (Patience, 2026). The investment has been accompanied by an intensive effort to frame AI adoption as an urgent educational priority.

Policymakers have shown a prescriptive optimism toward the adoption of AIED despite a lack of scientific evidence or grounding in educational theory (Ro, 2025). The April 2026 U.S. Department of Education final priority makes AI literacy, educator AI training, and AI integration eligible for favorable treatment across all discretionary grant competitions (U.S. Department of Education, 2026). In highly competitive grantmaking favorable treatment can function as a practical condition of competitiveness. This financial pressure normalizes adoption as a foregone conclusion, a stance epitomized by the New York City Department of Education's declaration that "the question is not whether A.I. belongs in schools," but how to govern it (Winter, 2026). The conclusion has already been determined – hence the title of Weidlich et al.'s (2025) article "ChatGPT in Education: An Effect in Search of a Cause."

1.2 The demand for evidence

At the same time a backlash to AIED is forming. Wallace-Wells (2026) writes in *The New York Times* about AI tools for education, "to date there is pretty limited evidence that they might help"

(para. 13). The financial stakes create intense demand for evidence that AI works: schools need guidance; policymakers need justification; corporations need legitimacy; shareholders need signs of progress; and skepticism requires rebuttal.

The effects of this pressure also reach academic literature. New studies are published regularly concluding they have provided an empirical basis for widescale adoption of AI in schools. Many of these studies overgeneralize, such as an article that characterizes Harvard undergraduates enrolled in physics classes as “broadly representative of those found across a range of institutions” (Kestin et al., 2025, p. 2). The study only assesses short-term learning, another common limitation of AIED studies and one which is problematic given Bastani et al.’s (2025) finding that AI intervention improves math scores in the short-term but harms long-term retention. Despite these limitations, Kestin et al. (2025) conclude their study is “a compelling case for [AI’s] broad adoption in learning environments” (p. 1).

A retracted meta-analysis made similarly bold claims: “The broad use of ChatGPT at various grade levels and in different types of courses should be encouraged to support diverse learning needs... Continuous use of ChatGPT should be ensured to support student learning” (Wang & Fan, 2025, p. 1). It attracted attention and citations, despite being identified as methodologically faulty within a week of its publication (Bartoš et al., 2025; Ingebrigtsen & Lukic, 2025) and being retracted on April 22, 2026 (notably after it had already been analyzed for this paper). It took nearly a year for the article to be retracted, during which time it accrued more than 500 citations (Boyes, 2026). It is likely to accumulate many more citations as manuscripts already in process at the time of its retraction are published. This paper examines how meta-analysis, the technique used by Wang and Fan (2025), has been misapplied to manufacture unwarranted certainty in the service of AIED advocacy.

Meta-analyses are especially consequential in this setting because they appear to offer an authoritative answer for policymakers (Tuomi, 2025). They present easy-to-understand, concrete, and quantitative conclusions, backed by complicated statistics and diagrams impenetrable to all but experts.

Recently, there has been an explosion of meta-analyses making broad claims about AI improving education. The claim itself is problematic: AI is a tool, not a coherent intervention. Weidlich et al. (2025) write, “speaking of the learning effects of ChatGPT amounts to a non sequitur; learning does not per se follow from a tool with an entirely different purpose” (p. 4).

1.3 Study purpose and analytic framework

This study examines how broad claims about AI’s educational effectiveness are constructed, evaluated, and circulated through recent AIED meta-analyses. The analysis focuses on four connected levels of validity. I evaluate the coherence of their constructs, treatment of dependent effects and heterogeneity, assessment of publication bias, and accuracy of the primary studies and effect sizes on which their conclusions depend. I also examine subsequent uptake of these meta-analyses, considering both the extent to which they have been cited and whether citing authors critically engage with their methodological limitations or instead reproduce their conclusions as established evidence.

2. The technique of meta-analysis

2.1 Effective meta-analysis design

Meta-analysis is a form of synthetic research that combines data from a set of studies on the same topic to create a single summary metric quantifying the direction and magnitude of an effect. A critical requirement for meta-analysis is that the effect being synthesized is coherent. Meta-analysis can tolerate variation in how primary studies are operationalized, but only if they share a common conceptual end (Higgins et al., 2019). This requires both that the independent variable represents a common intervention construct, and that the dependent variable represents a common outcome construct. Attempting a meta-analysis of studies measuring different things obscures their meaningful differences under a meaningless average (Lipsey & Wilson, 2001). The effect sizes extracted from primary studies must be interpretable as estimates of the same intervention-outcome relationship, or at minimum a family of similar relationships.

2.2 Heterogeneity

Heterogeneity is the degree to which the study results differ beyond what would be expected from random sampling error alone. The I^2 statistic estimates the proportion of variation amongst primary studies attributable to true differences. Because I^2 is a relative measure and does not indicate the magnitude of those differences, random effects meta-analyses should also report the estimated between-study variance, τ^2 (Borenstein et al., 2017; Higgins et al., 2019; Wu et al., 2025). When heterogeneity is substantial, the prediction interval should be reported prominently and interpreted alongside the pooled mean (Higgins et al., 2019, p. 263; IntHout et al., 2016; Tipton et al., 2023). Additional explanation of these statistics is provided in Appendix F.

Bryan et al. (2021) call for a “heterogeneity revolution” in which variation is treated as a central object of analysis providing evidence about when and for whom interventions work. This can be accomplished via moderator analysis: a description of moderator analysis can be found in Appendix F. Tipton et al. (2023) advise against testing moderators one at a time by repeatedly dividing the evidence into subgroups. Such analyses discard information, have low statistical power, and cannot distinguish among correlated study characteristics. They recommend moderators be tested jointly by multivariable meta-regression. Continuous moderators should remain continuous rather than be divided at arbitrary cut points. Statistical testing should account for dependence when individual studies contribute multiple effects (Becker et al., 2004; Tipton et al., 2019a, 2019b; Wu et al., 2025).

Testing multiple moderators increases the probability of significant results arising by chance. *The Cochrane Handbook* recommends adjusting the criterion for statistical significance when more than one or two study characteristics are investigated (Higgins et al., 2019, p. 269). For the present audit, studies were flagged for testing more than five moderator variables without adjusting the critical value. This deliberately permissive threshold is an operational screening criterion, not an assertion that five unadjusted moderator tests are generally safe. An exploratory analysis may reasonably include a larger

number of unadjusted moderator tests, provided results are clearly identified as exploratory and interpreted cautiously. Moderator analysis requires a considerable sample size to be effective. Based on the work of Cuijpers et al. (2021) this audit flags studies with subgroups of fewer than 10 studies as likely underpowered. See Appendix F for a detailed account of how this criterion was determined.

2.3 Publication bias

Studies with positive results are more likely to be published (Rosenthal, 1979), especially studies applying behavioral methodologies on humans (Fanelli, 2010). The framing of the field of AIED further strengthens confirmatory bias. A study of calls for papers in AIED found two-thirds (65.8%) specifically invited research on the benefits of AIED, seemingly disinviting null or negative results (O'Neill, in review). This makes publication bias a predictable threat to the validity of AIED meta-analyses.

Multiple methods exist to quantify publication bias: see Appendix G for details on such methods. Regardless of the methods used, it is impossible to prove publication bias has no effect on a meta-analysis (Sinha & Montori, 2006; Vevea & Woods, 2005). The best available quantitative methods (e.g. Egger's test) have very low statistical power (Lin et al., 2018; Sutton et al., 2000), and a perfectly symmetrical funnel plot—or a nonsignificant Egger's test—does not fully rule out publication bias.

Because publication bias is a well-established phenomenon and quantification methods cannot conclusively exclude it, meta-analyses should correct for it regardless of its detection (Ropovik et al., 2021; Wu et al., 2025). This concern is especially relevant in AIED, which has been documented to have strong bias towards positive results (Higgins et al., 2012; O'Neill, in review; Ro, 2025; Tamim et al., 2021). All methods of correcting for publication bias are imperfect, as they require assumptions about the nature of missing data that can never be tested (Copas & Shi, 2000). Several methods are described in Appendix H. Wu et al. (2025, p. 147) recommend selecting two or more approaches for correcting for publication bias from recent research.

3. Methods

3.1 Research design: Forensic methodological audit

This study uses a forensic methodological audit to examine recent broad-claim meta-analyses of AIED and a sample of the primary studies on which they relied.

3.2 Meta-analysis selection

The search was designed to produce a systematic, reproducible corpus; exact search strings are in Appendix A. Articles were only considered from peer-reviewed journals published after January 1, 2023, after the release of ChatGPT, with abstracts that claim that AI or a specific AI technology impacted education, academic performance, academic achievement, learning outcomes, learning performance, or higher-order thinking. Studies that narrowed to specific skills (e.g. computer programming, speaking a new language), sectors (e.g. health professions training), or populations (e.g. students with autism spectrum disorder) were excluded. This resulted in 14 remaining meta-analyses, listed in Appendix B.

3.3 Structured document analysis: AIED meta-analysis audit procedure

The criterion used in assessing meta-analyses are rationalized in greater detail in Appendix I. The following assessments were recorded for each meta-analysis:

3.3.1 Construct validity

- Record primary claim.
- Identify independent variable and dependent variable. Evaluate their coherence.
- Record inclusion and exclusion criteria used in study selection.

3.3.2 Statistical independence of effect sizes

- Are dependent effect sizes present? How are they addressed?

3.3.3 Heterogeneity

- Is I^2 reported and accurately interpreted? Is τ^2 reported and accurately interpreted?
- If I^2 exceeds 50%: is a prediction interval reported and substantively interpreted?

- How many moderator variables are analyzed?
 - Are moderator variables theoretically rationalized?
 - Do moderator analyses violate statistical independence?
 - What is the smallest subgroup size?
 - Is confounding risk minimized (e.g. by multivariable regression)?
 - If >5 moderators: is a sensitivity adjustment applied to mitigate type I error inflation?

3.3.4 Publication Bias

- What tests are utilized (e.g. visual inspection of funnel plots, Egger's linear regression test, etc.)?
- Are any statistical requirements for independence met?
- Are statistics properly reported? Are low-powered tests of publication bias overinterpreted?
- What adjustments for publication bias were applied (e.g. selection models, trim-and-fill)?

3.4 Selection of primary studies for analysis

Five primary studies were randomly selected from each of 12 of the 14 meta-analyses; one meta-analysis predominantly included studies in Korean (Lee & Kwon, 2024) and another did not identify the studies included (Zheng et al., 2023). Because some primary studies appeared in multiple meta-analyses, the 60 random selections represented 46 unique publications for vetting. Additional primary studies were selected when there was a priori reason to suspect misalignment with the meta-analysis claim, such as an apparent title mismatch or correlational design. Ultimately, 13 publications were targeted for vetting in addition to the 46 randomly selected publications for a total of 59 vetted primary studies. Prevalence estimates are reported using only the randomly selected studies.

3.5 Evidentiary audit of primary studies

3.5.1 Construct alignment

- Does the experiment/intervention fit the inclusion criteria of the meta-analysis?
- Do the dependent and independent variables match the claim in the meta-analysis?

3.5.2 Causal design

- Is there a control group that does not interact with AI?
- Are groups randomized?
- Is the control group otherwise equivalent to the experimental/intervention groups?

3.5.3 Data reliability and transparency

- Is the sample size extremely small ($N \leq 30$)?
- Is the basis for the effect size extracted by the meta-analysis clear?
- Are there signs of fabricated, unverifiable, or incoherent scholarship?

3.6 Targeted citation uptake analysis

Two meta-analyses were selected to analyze how they have been cited by subsequent literature. Lee and Kwon (2024) were selected because their writing is susceptible to misinterpretation: although they declare a simple “effect of AI education” (p22859) without specifying a dependent variable, their study actually examines teaching about AI, rather than using AI itself. Wang and Fan (2025) were selected because the article had attracted substantial citation attention and has now been retracted, making it the clearest case of invalid claims spreading.

For each selected meta-analysis, citing articles were identified through ResearchGate because it provides excerpts showing the context in which a source is cited. ResearchGate provided excerpts for 14 citations to Lee and Kwon (2024); all 14 were analyzed. For Wang and Fan (2025), the 20 most recent citations with available excerpts as of May 31, 2026, were analyzed.

4. Findings

4.1 Coherence of meta-analysis claims

Across examined meta-analyses, the most basic threat to interpretability was not statistical but conceptual. All but two defined the treatment as a tool such as ChatGPT or GenAI (see Table 1, p. 11). A product name is not a pedagogy. Treating exposure to ChatGPT as a common intervention is akin to

Table 1*Evaluating Construct Coherence in Meta-Analyses*

Study	Claim	IV Coherence	DV Coherence	Notes
Dai et al. (2024)	AI-powered virtual agents improve learning in computer simulations.	Strong	Very poor	AI-powered agents clearly defined. Learning includes behavioral, cognitive, and social outcomes.
Doo & Park (2026)	ChatGPT improves learning achievement.	Very poor	Incoherent	Uses tool as IV. Achievement includes metacognition, behaviors, and attitudes.
Gu & Yan (2025)	GenAI increases academic achievement.	Incoherent	Poor	Uses tool as IV; GenAI is inclusive of LLMs, image, and video production. Academic achievement is multi-faceted but excludes perception and creativity.
Lee & Kwon (2024)	AI education has a positive effect size.	Ambiguous	Ambiguous	No clear DV in abstract or conclusion. Claims are highly misleading.
Li et al. (2025)	AI technologies improve educational outcomes in personalized K–12 STEM education.	Very poor	Incoherent	Uses tool as IV. Operationalizes DV via knowledge, skills, engagement, attitude, and perceptions.
Liu et al. (2025)	ChatGPT increases academic achievement.	Very poor	Incoherent	Uses tool as IV. Academic achievement includes knowledge, skills, motivation, attitude, creativity, and more.
Ma & Zhong (2025)	GenAI has a positive impact on learning outcomes.	Very poor	Incoherent	Uses tool as IV. Learning outcomes include cognitive, affective, and competency in a single variable.
Sun & Zhou (2024)	GenAI improves college students' academic achievement.	Very poor	Incoherent	Uses tool as IV. Academic achievement includes knowledge, skills, motivation, attitude, creativity, and more.
Wang & Fan (2025)	ChatGPT has a positive effect on learning performance.	Very poor	Ambiguous	Uses tool as IV. Learning performance is never defined.
Wu & Yu (2024)	AI chatbots have a large effect on learning outcomes.	Very poor	Incoherent	Uses tool as IV. Learning outcomes is performance, motivation, self-efficacy, anxiety, and perceived learning.
Wu et al. (2026)	ChatGPT improves learning outcomes.	Very poor	Incoherent	Uses tool as IV. Learning outcomes broadly defined.
Zhao et al. (2025)	GenAI increases higher-order thinking skills.	Very poor	Poor	Uses tool as IV. Higher order thinking skills is somewhat multi-faceted.
Zheng et al. (2023)	AI increases learning achievement and learning perception.	Incoherent	Poor	AI is so broadly defined that many dissimilar tools fit. DVs are both somewhat multi-faceted.
Zhu et al. (2025)	GenAI improves learning outcomes.	Incoherent	Incoherent	GenAI defined by indefinite list of examples; DV combines intellectual, social-emotional, and physical.

Note. IV = independent variable; DV = dependent variable. Ratings: strong = aligned with claim; poor = serious construct drift; very poor = unstable but identifiable construct; incoherent = no stable pooled referent; ambiguous = unclear or undefined.

attempting a meta-analysis on the effects of paper: the medium may be used in many instructional designs but is not itself the instructional design. The relevant treatment is the designed learning activity in which the medium is used, not the medium alone (Tuomi, 2025; Weidlich et al., 2025).

Outcomes such as “academic achievement” are often explicitly defined as multidimensional concepts that can include test scores, homework quality, perceptions of learning, critical thinking, problem solving, motivation, self-efficacy, attitudes, and engagement. This is a fundamental misunderstanding of the limitations of meta-analysis. Naming a construct multidimensional does not make its dimensions interchangeable. Pooling these outcomes produces an average of unlike effects that may be mathematically calculable but lacks a stable substantive meaning (Lipsey & Wilson, 2001).

Population and setting are also key factors to consider in determining the coherence of a meta-analysis (Dekkers, 2018). Of the 14 meta-analyses, 11 have no limitation on population and nearly all draw studies internationally. Educational populations and settings are not interchangeable because learning is shaped by learner development, prior knowledge, culture, and instructional environment (National Academies of Sciences, Engineering, and Medicine, 2018). In most of the assessed meta-analyses, “students” became an all-purpose category spanning children to medical trainees. Such diversity could potentially yield additional insight if variation across populations and settings were treated as a substantive object of analysis rather than averaged away (Conaway et al., 2024). Unfortunately, none of the meta-analyses included enough independent studies to support meaningful moderator analysis (see Section 4.4).

4.2 Audit of primary studies

A total of 59 primary studies were chosen for vetting: 46 via random selection and 13 via targeted selection (see section 3.4 for details). The audit of the randomly selected studies found more than half presented validity concerns for the meta-analysis ($n = 28$, 61%). Targeted studies were all problematic. Primary studies that are highly effective research may still be flagged if they were misused

by the meta-analysis. The most common validity concerns were dependent variable mismatch ($n = 15$), independent variable mismatch ($n = 11$), experimental design ($n = 7$), data extraction ($n = 6$), no control group ($n = 6$), and nonrandom group assignments ($n = 6$). These counts are not mutually exclusive as some vetted primary studies did not yield any concerns and others yielded concerns in multiple categories. Examples are described below; more detailed lists can be found in appendices C and D.

4.2.1 Dependent variable mismatch

Fifteen audited primary studies measured outcomes that did not validly represent the claim of the meta-analysis: Yilmaz and Karaoglan Yilmaz (2023) measures student perception of learning, motivation, and self-efficacy through Likert scales, but six meta-analyses use it as evidence of learning. Studies of mental health interventions that measure anxiety and depression are used as evidence of improved learning outcomes (Beidel et al., 2021; Klos et al., 2021; Liu et al., 2022).

4.2.2 Independent variable mismatch or independent variable not isolated

Eleven audited primary studies' interventions did not match the intervention construct, and an additional five did not isolate the independent variable making it impossible to attribute the effects of the study to AI. Bachiri et al. (2023) compare gamification with automatic question generation to traditional assessment. It is included by three meta-analyses focused on GenAI when it seems likely that the gamification was the most consequential part of the intervention.

4.2.3 Data extraction problems

Six audited primary studies did not present enough data to be transparent about their process, had errors in statistics, or had an extracted effect size used in a meta-analysis that was not easily traced or replicated. Yang et al. (2025) was extracted as a negative effect size by Wang and Fan (2025) but should be a positive effect size. Sun and Zhou (2024) extract a large effect size from Sadek and Mohamed (2023) which I cannot replicate ($g \approx 2.05$ as read from their Forest plot). Using the reported means and standard deviations I calculated $g = 0.72$ (see Appendix E).

4.2.4 Other experimental design concerns

Seven audited primary studies included experimental design problems such as interventions for the experimental group that didn't align with the inclusion criteria for the meta-analysis; six more did not randomize groups; six lacked a control group; and one had the control group utilizing AI.

4.2.5 Incoherent and unreliable studies

Ji et al. (2023) presents data in a misleading way: tables show number of participants ($n = 21$ and $n = 15$), but the unit of analysis is teams ($n = 5$ and $n = 3$). I verified Ji et al.'s statistics are correctly calculated, but the presentation of more palatable sample sizes confused meta-analysts. Both Liu et al. (2025) and Sun and Zhou (2024) included the Ji et al. (2023) as a primary study in their meta-analysis, and both extracted an effect size based on participants ($n = 21$ and $n = 15$) instead of teams, which are the proper unit of analysis ($n = 5$ and $n = 3$), thereby miscalculating effect size.

Alneyadi and Wardat (2023) include references to articles about ChatGPT dated prior to the existence of ChatGPT such as "Bétrancourt, M. (2019d). The potential of ChatGPTs in teaching and learning. *Computers in Human Behavior*, 92, 182-186." This is followed by an incorrect DOI that loads an unrelated article. There are six such references; four are attributed to Mirielle Bétrancourt, but no matching works are found on her CV. The study is itself incoherent. For example, it reports $\eta^2 = .0403$, which it states is "greater than 0.14." (p. 9) This study is included in six meta-analyses. Alneyadi and Wardat (2024) suffers similar problems. There are five references to articles about ChatGPT dated prior to the existence of ChatGPT. In table 3, η^2 is reported as 1.374, but η^2 cannot exceed 1. In several rows of the table mean squares do not equal the sum of squares divided by degrees of freedom, which should always be true for MANCOVA. This study was included in three meta-analyses.

4.3 Statistical independence

Twelve meta-analyses treated effect sizes drawn from the same primary study as independent, violating the assumptions of some or all of their statistics (see Table 2, p. 15).

Table 2*Treatment of Dependent Effect Sizes in Independence-Sensitive Analyses*

Study	Dependent Effect Sizes?	Strategy	Effect Size	I^2	Mod.	Bias
Dai et al. (2024)	Yes	Robust variance estimation accounts for dependent effects in overall effect.	OK	No	No	No
Doo & Park (2026)	Yes	Treats all as independent. "... a total of 86 independent studies in 22 publications..." (p270)	No	No	No	No
Gu & Yan (2025)	Yes	"... we ignored independence because only a very small number of studies included more than one effect size." (p13)	No	No	No	No
Lee & Kwon (2024)	Yes	Use publications as unit of analysis except in moderator analysis.	OK	OK	No	No
Li et al. (2025)	Yes	"Given the limited number of qualified articles... the aggregation of diverse outcomes from the same studies was not pursued." (p10)	No	No	No	No
Liu et al. (2025)	No	Unnecessary; only one effect size extracted from each publication.	OK	OK	OK	—
Ma & Zhong (2025)	Yes	Acknowledge in limitations that dependence was not addressed.	No	No	No	No
Sun & Zhou (2024)	Yes	Specifically refer to the set of extracted effect sizes as "65 independent studies." (p22)	No	No	No	No
Wang & Fan (2025)	Yes	Unsuccessful; stated strategy would be OK for effect size at least, but not correctly executed.	No	No	No	—
Wu & Yu (2024)	Yes	Use only one effect per study per outcome, but then pool multiple outcomes for statistics.	No	No	No	No
Wu et al. (2026)	Yes	Transparency problem: no strategy discussed; degrees of freedom suggest aggregation for pooled effect and I^2 .	OK	OK	No	No
Zhao et al. (2025)	Yes	Dependence not addressed	No	No	No	No
Zheng et al. (2023)	No	Unnecessary; only one effect size extracted from each publication.	OK	—	OK	—
Zhu et al. (2025)	Yes	Dependence not addressed	No	No	No	No

Note. "Dependent effect sizes?" indicates whether multiple effect sizes were extracted from the same primary study. Independence-sensitive analyses include the pooled effect size, I^2 , moderator analyses, and Egger's or Begg's tests for publication bias. Judgments were based on reported sample sizes, degrees of freedom, and funnel-plot points when necessary. Effect size = pooled effect size for the full meta-analysis; Mod. = moderator analysis; Bias = Egger's and/or Begg's test. OK = independence requirement met; No = independence requirement not met; — = analysis not conducted.

4.4 Heterogeneity

Reported heterogeneity was high in every meta-analysis that provided I^2 (range = 77.2%–94.4%); 12 of the 13 reported values exceeded 80%. Because I^2 is precision-dependent, these statistics do not by themselves indicate the absolute dispersion of true effects. Only four meta-analyses reported global τ^2 , and only two reported prediction intervals; both prediction intervals included zero. See Appendix F for a discussion of τ^2 , random-effects models, and moderator subgroup residual heterogeneity.

All 14 meta-analyses included moderator analysis; 12 violated the independence assumptions required by the statistical tests they employed. Moderator analyses were also severely underpowered. No meta-analysis met the baseline recommendation that the smallest subgroup include at least ten studies; given the extreme heterogeneity in this corpus, even larger subgroup sizes were actually needed (Cuijpers et al., 2021). Five meta-analyses included subgroups with only a single study, and three others had subgroups as small as two studies. A “moderator” represented by one or two studies cannot explain heterogeneity in any meaningful sense. At that point, the analysis is not evaluating a subgroup but describing isolated studies as if they represented broader categories.

Moderator variables were also generally examined separately, leaving correlated study characteristics unresolved. Educational level, subject, intervention duration, instructional design, and study quality are often confounded, yet no meta-analysis attempted to minimize confounding among moderators. Finally, eight meta-analyses evaluated more than five moderator variables, with some testing as many as 13, but none applied a more stringent critical value or another adjustment for Type I error inflation. Numerous unadjusted tests may be defensible in an explicitly exploratory analysis, but all 14 meta-analyses interpreted their moderator analysis as confirmatory and definitive.

4.5 Publication Bias

All meta-analyses attempted to assess publication bias (see Table 4, p. 18). The most common technique was to make a funnel plot. Egger’s test was used by 10 studies and one used only Begg’s test;

Table 3*Analysis of Heterogeneity in AIED Meta-Analyses*

Study	Global k (ES)	Global I^2 (%)	Global τ^2	Pred Int	No. Mods	Mod Ind	Mod Conf	α adj	Min Sub
Dai et al. (2024)	22 (49)	83.0%	.25	No	9	No	No	No	1
Doo & Park (2026)	22 (86)	83.4%	NR	Yes	4	No	No	—	2
Gu & Yan (2025)	19 (24)	94.4%	1.27	Yes	8	No	No	No	4
Lee & Kwon (2024)	64 (196)	85.0%	NR	No	8	No	No	No	7
Li et al. (2025)	32 (99)	89.7%	NR	No	6	No	No	No	1
Liu et al. (2025)	37 (37)	93.2%	.255	No	9	—	No	No	1
Ma & Zhong (2025)	34 (100)	91.2%	NR	No	5	No	No	No	9
Sun & Zhou (2024)	28 (65)	83.3%	.212	No	6	No	No	No	2
Wang & Fan (2025)	51 (72)	89.2%	NR	No	7	No	No	No	2
Wu & Yu (2024)	24 (42)	94.0%	NR	No	3	No	No	—	3
Wu et al. (2026)	35 (134)	91.4%	NR	No	5	No	No	—	1
Zhao et al. (2025)	29 (59)	77.2%	NR	No	4	No	No	—	6
Zheng et al. (2023)	24 (24)	NR	NR	No	13	—	No	No	1
Zhu et al. (2025)	26 (44)	82.8%	NR	No	5	No	No	—	4

Note. k = number of primary studies synthesized; ES = total number of unique effect sizes extracted; Pred Int = reporting of prediction interval; No. Mods = number of moderator variables evaluated; Mod Ind = statistical tests for moderator groups met all necessary assumptions of independence; Mod Conf = makes attempt to address confounding due to collinearity; α adj = lowered critical value due to large number of moderating variables; Min Sub ($k_{\min}$) = smallest subgroup in moderator analysis; NR = not reported; — = not applicable.

Table 4*Publication Bias in AIED Meta-Analyses*

Study	Tests of Publication Bias	Additional Concerns	Data Given	Ind	Bias Adj	Report Adj	Valid
Dai et al. (2024)	Egger's test; trim-and-fill	Egger $p < .10$ treated as nonsignificant.	No	No	Yes	No	No
Doo & Park (2026)	Funnel plot; Egger's test	None	No	No	No	No	No
Gu & Yan (2025)	Funnel plot; Egger's test; Vevea & Woods model	Insufficient info on Vevea & Woods selection model	No	No	Yes	No	No
Lee & Kwon (2024)	Funnel plot; Egger's test; fail-safe N metrics; publication subgroup analysis	No statistics reported for Egger's test	No	No	No	No	No
Li et al. (2025)	Fail-safe N metrics	None	Yes	—	No	No	No
Liu et al. (2025)	Funnel plot; Egger's test; Kendall's tau rank test; fail-safe N metrics; trim-and-fill methods.	Begg's test is a modified Kendall's tau. This is either misnamed or misapplied.	No	Yes	Yes	No	No
Ma & Zhong (2025)	Funnel plot; Egger's test	No statistics reported for Egger's test	No	No	No	No	No
Sun & Zhou (2024)	Funnel plot; Egger's test	Egger $p < .10$ treated as nonsignificant.	No	No	No	No	No
Wang & Fan (2025)	Funnel plot; Rosenthal's fail-safe N	None	Yes	—	No	No	No
Wu & Yu (2024)	Funnel plot; Egger's test; Begg's test; Trim-and-fill	Applies tests to small subgroups (down to $n = 3$)	No	No	Yes	Yes	No
Wu et al. (2026)	Funnel plot; fail-safe N metrics; Begg's test	No rationale for selecting Begg's test	No	No	No	No	No
Zhao et al. (2025)	Funnel plot; Egger's test	Egger $p < .10$ treated as nonsignificant.	No	No	No	No	No
Zheng et al. (2023)	Funnel plot; fail-safe N metrics	None	Yes	—	No	No	No
Zhu et al. (2025)	Funnel plot; Egger's test; fail-safe N metrics	Egger $p < .10$ treated as nonsignificant.	No	No	No	No	No

Note. Data Given = all standard test statistics are reported (e.g., β_0 , t , and p for Egger's test); Ind = whether independence assumptions required for Egger's or Begg's tests were met; Bias Adj = whether a correction for publication bias was calculated; Report Adj = whether the adjusted effect size was reported in the abstract and conclusion; Valid = whether the meta-analysis provided a valid and sufficient accounting of publication bias. — = not applicable.

10 of those 11 analyses violated assumptions of independence by entering multiple effect sizes from the same paper as separate data points, an especially consequential failure given how many of the tests were borderline significant. The critical value for Egger's regression test was always set at $p < .05$, with 7 studies declaring no significance and 3 finding significance. Had the recommended critical value of $p < .10$ (Higgins et al., 2019; Jin et al., 2014; Peters et al., 2006; Sterne et al., 2011) been applied, five additional tests would have been classified as significant – and possibly Ma and Zhong (2025) who declared Egger's test nonsignificant without giving any statistics. Only Liu et al. (2025) made valid calculations assessing publication bias, as other studies failed to account for the dependence of effect sizes drawn from the same study or relied on discredited fail-safe N metrics.

Meta-analyses persistently failed the criterion of appropriately hedging inferences from low power tests. Negative Egger's tests were interpreted as proof of no publication bias, Wu and Yu (2024) write, "Egger's test indicates an absence of publication bias..." (p. 22). Sun and Zhou (2024) write Egger's test indicated "there was no significant publication bias... the results of the analyses in this study are relatively stable and reliable" (p. 16).

Widely discredited fail-safe N metrics (Heene, 2010; Scargle, 2000) were used in 6 of the meta-analyses. For 3 meta-analyses this was the only attempt to quantify a risk of publication bias. Interpretations of fail-safe N metrics treated them as authoritative. Li et al. (2025) conclude "that the impact of publication bias on the effect size in the current review was trivial" (p. 12). Lee and Kwon (2024) ascribed greater authority to fail-safe N metrics than Egger's test, using them to dismiss the significance of the more powerful Egger's test, "Since the coefficient (N) is 5,952, which is greater than $5k + 10$ (k : number of studies), the results of this study can be interpreted as being reliable" (p. 14).

Wu et al. (2026) use Begg's test instead of the preferred Egger's test without providing a rationalization. The use of the substantially lower-powered Begg test is not justified. Its $p = .079$ result was misinterpreted as nonsignificant, when $p < .10$ should be the critical value for low-powered tests of

funnel plot symmetry. They also run two widely discredited fail-safe N metrics and interpret the findings as proof of validity. "In summary, these three methods showed no significant publication bias, and the effect sizes were stable. The conclusions accurately reflect the actual situation" (Wu et al., 2026, p. 32).

Wu and Yu (2024) come the closest to an effective treatment by acknowledging publication bias and adjusting via trim-and-fill method, using the adjusted value as the primary claim, but they fail to report the original unadjusted estimate making it impossible to consider the impact of their adjustment. Three other meta-analyses calculated an adjustment for publication bias, but then used the original unadjusted effect size as the primary finding stated in the abstract and conclusion.

The importance of publication-bias adjustment is illustrated by Wang and Fan (2025). Their meta-analysis – now retracted – reported a large positive effect of ChatGPT on learning performance and made broad implementation recommendations. They dismissed the possibility of publication bias based on fail-safe N metrics. Bartoš et al. (2025) reanalyzed the data using multiple methods of publication-bias adjustment and found the apparent effect disappeared entirely under all models.

4.6 Targeted citation uptake analysis

Lee and Kwon (2024) examine if teaching students about AI actually yields gains in knowledge about AI, but the claim stated in their abstract is "The overall effect size of AI education is 0.897 (95%CI = 0.781 ~ 1.013, $p < .05$), which can be interpreted as a large effect." (p. 22859). This ambiguous claim invites the misinterpretation that Lee and Kwon (2024) studied application of AI to general education. An analysis of papers that cited Lee and Kwon (2024) was conducted. The majority of the 14 papers cited Lee and Kwon (2024) incorrectly (see Table 5, p. 21), but consistent with the misleading wording of the abstract and conclusion: "the effect size of AI education" was interpreted to mean the impact of AI technology on educational outcomes ($n = 8$ papers, 57%). Surprisingly, most other papers citing Lee and Kwon (2024) also did so incorrectly ($n = 4$ papers, 29%). One paper cites Lee and Kwon (2024) as support for improved efficiency after adopting automatic grading tools; another for defining AI as being able to

Table 5

Assessing a Sample of Papers Citing Lee and Kwon (2024) and Wang and Fan (2025)

Lee and Kwon (2024) Citations	n	%
Incorrectly cited as evidence that the integration of AI improves learning.	8	57
Other incorrect citation	4	29
Appropriate citation	2	14
Wang and Fan (2025) Post-Retraction Citations	n	%
Cited as authoritative evidence that the integration of AI improves learning.	12	60
Partially critical citation	2	10
Critical citation	1	5
Incorrectly cited	5	25
Acknowledged Wang & Fan (2025) retraction	0	0

Note. Percentages are rounded to the nearest whole number.

“make decisions about how to teach students on its own,” neither of which appears in the paper.

Appropriate citations were uncommon ($n = 2$ papers, 14%).

All 20 papers citing Wang and Fan (2025) were published after the April 22, 2026 retraction; none acknowledged the retraction. This suggests that retraction is unlikely to influence papers already in the review process. Most ($n = 12$ papers, 60%) cited Wang and Fan (2025) as authoritative support that ChatGPT drives large performance gains in education. Two papers cited Wang and Fan (2025) in a partly critical way, for example accepting it was strong support for short-term learning benefits while questioning if interventions might still undermine deep learning ($n = 2$ papers, 10%); another paper was explicitly critical ($n = 1$ paper, 5%). The remaining citations were incorrect ($n = 5$ papers, 25%), for example citing Wang and Fan (2025) as determining which characteristics of AIED interventions make AIED more likely to be adopted.

5. Discussion

5.1 Methodological failure is chronic in published AIED meta-analyses

Of the 14 meta-analyses audited, none had a coherent construct. A majority of the primary studies randomly chosen for vetting were invalid or misaligned. Twelve of the 14 meta-analyses failed to address dependence among effect sizes drawn from the same study. Only two included a prediction interval, which is necessary to intelligibly interpret a meta-analysis with high heterogeneity. More than half drew conclusions using subgroups as small as $k = 1$ or $k = 2$. None provided a valid and sufficient assessment of publication bias. The meta-analyses audited had too many technical errors to catalog more than a slight fraction in a single journal article: see Appendix C for a critique of each meta-analysis.

5.2 Comparison to meta-analyses on other topics

Meta-analyses across fields practice poor methodology. Tipton et al. (2019b) report that 83% do not control for confounding, and 88% do not use multiple comparison corrections, despite a consensus that these are necessary practices. Wu et al. (2025) analyzed the first 100 meta-analyses published in November 2021 in ten disciplines, including education, finding several similarities to the current corpus of AIED meta-analyses including high heterogeneity and failure to report τ^2 . Wu et al. found a median I^2 value of 86% in education meta-analyses with 35% of meta-analyses reporting a value for τ^2 ; in this study I^2 had median value of 89.2% and 29% of meta-analyses reported a value for τ^2 .

In other ways, this corpus of AIED meta-analyses are more problematic than typical education meta-analyses including failure to adjust for dependent effect sizes and addressing potential confounding among moderator variables. Wu et al. (2025) found 23% of education meta-analyses had dependent effect sizes but ignored them, compared with 64% in the present AIED corpus. Wu et al. also found that 81% of education meta-analyses used meta-regression, but only 32% used multivariable meta-regression necessary to control for confounding; in the current study 14% used meta-regression, with 0% using multivariable meta-regression.

Ignoring dependent effect sizes is an inferential problem impacting heterogeneity statistics, moderator analyses, and publication-bias tests. Treating dependent effects as independent inflates the apparent amount of evidence, underestimates standard errors, narrows confidence intervals, and makes statistical significance easier to obtain. It can also distort the pooled mean by giving disproportionate weight to studies that contribute many correlated outcomes (Becker et al., 2004; Hedges et al., 2010). It is also easily avoided; there are techniques available that can utilize all relevant effect sizes and account for dependence that have now been integrated into many software packages (Tipton et al., 2019a).

5.3 Beyond statistical repair: Primary study selection and ecological validity

The most conspicuous errors in these meta-analyses are statistical because they can be stated as identifiable violations: effect sizes were extracted incorrectly, dependent estimates were treated as independent, and methods to detect publication bias were misapplied then misinterpreted. These errors can be corrected by reanalyzing the same evidence. More fundamental problems arise when the evidence itself is unsuitable. Statistical adjustment cannot repair a synthesis built from invalid or construct-misaligned primary studies.

Five meta-analyses included a formal metric for evaluating primary study quality. A primary study with fabricated references and impossible values for statistics was included by six meta-analyses (Alneyadi & Wardat, 2023), including four of the five that had utilized metrics to evaluate primary study quality. An inclusion error of this severity not only adds a contaminated effect size to the meta-analysis but also calls into question the process by which primary studies were selected. The large proportion of problematic primary studies confirms poor selection procedures are widespread: 61% of the randomly chosen primary studies were methodologically deficient or construct-misaligned to the meta-analysis.

Even valid primary studies would be incapable of supporting the broader claims made by these meta-analyses. The pooled estimates frequently crossed substantial differences in instructional purpose, learner population, setting, duration, implementation, and outcome. As a result, the synthesis may

estimate an average across studies without identifying any coherent educational condition to which that average applies. This is not a problem that can be resolved by recomputing an effect size; it concerns whether the synthesized evidence answers a meaningful question about educational practice at all.

5.4 A cascade of errors results in a concrete policy recommendation

Several meta-analyses include sample size as a moderating variable. Study sample size is not class size and cannot be interpreted as an instructional feature. Liu et al. (2025) and Sun and Zhou (2024) nevertheless interpret it as class size, although in many of the primary studies the entire sample did not participate in the treatment as a single class. Both meta-analyses have a small subgroup of studies with sample sizes of 21–40, which is dominated by the extreme effect size of the Ji et al. (2023) primary study. Ji et al. (2023) has misleading tables that list the number of students present as though they are sample sizes ($N = 36$), but groups of students are the unit of analysis ($N = 8$). As a result effect size is miscalculated by Liu et al. (2025) and Sun and Zhou (2024) and also placed into the incorrect category for sample size. That misplacement directly drives the finding in both meta-analyses that 21–40 students is the ideal intervention size for ChatGPT/GenAI. They then theorize why it makes sense and recommend ChatGPT/GenAI interventions be designed for this number of students.

The sample size variable is fundamentally misunderstood. Primary studies are incorrectly assigned to subgroups. The calculations are incorrect. Subgroup sizes are too small for valid results. The studies draw conclusions and make policy recommendations from this analysis that layered one error on top of another. Despite these errors, both studies' conclusions are published in prestigious journals and have accumulated many citations.

5.5 The circulation of flawed evidence

The failures identified in this review are not found in obscure or marginal publications. The examined meta-analyses appeared in journals published by Springer Nature, Wiley, SAGE, Taylor & Francis, and MDPI (see Table 6, p. 25). Several were published in journals with substantial impact

Table 6*Publication Venues and Scholarly Reach of Examined Meta-Analyses*

Study	Journal	Publisher	2024 JIF	Citations
Dai et al. (2024)	Educational Psychology Review	Springer Nature	8.8	126
Doo & Park (2026)	International Review of Research in Open and Distributed Learning	Athabasca University	1.9	5
Gu & Yan (2025)	Journal of Educational Computing Research	SAGE Publications	4.9	34
Lee & Kwon (2024)	Education and Information Technologies	Springer Nature	5.4	34
Li et al. (2025)	International Journal of STEM Education	Springer Nature	8.0	32
Liu et al. (2025)	Journal of Computer Assisted Learning	Wiley	4.6	44
Ma & Zhong (2025)	Journal of Computer Assisted Learning	Wiley	4.6	41
Sun & Zhou (2024)	Journal of Educational Computing Research	SAGE Publications	4.9	175
Wang & Fan (2025)	Humanities and Social Sciences Communications	Springer Nature	3.6	712
Wu & Yu (2024)	British Journal of Educational Technology	Wiley	8.1	817
Wu et al. (2026)	Humanities and Social Sciences Communications	Springer Nature	3.6	12
Zhao et al. (2025)	Journal of Intelligence	MDPI	3.4	29
Zheng et al. (2023)	Interactive Learning Environments	Taylor & Francis	5.3	294
Zhu et al. (2025)	Education and Information Technologies	Springer Nature	5.4	112

Note. JIF = Journal Impact Factor. Citation counts were recorded from Google Scholar on July 25, 2026.

factors, including *Educational Psychology Review*, *British Journal of Educational Technology*, and *International Journal of STEM Education*. By July 25, 2026, the examined articles had accumulated 2,467 combined citations. This reach is especially consequential because section 4.6 (above) showed AIED meta-analyses are often cited uncritically as authoritative evidence.

Publication of flawed meta-analyses does more than circulate unsupported conclusions; it also normalizes the methods that produced them. Researchers conducting new meta-analyses may reasonably look to recently published, peer-reviewed syntheses in prestigious journals as models for acceptable practice. When those models repeatedly rely on weak or inappropriate procedures, the problem becomes self-reinforcing. For example, the frequent use of fail-safe *N* metrics across the examined corpus gives a discredited method the appearance of legitimate practice. Publication also confers methodological authority on the authors themselves, making researchers who have demonstrated serious misunderstandings of meta-analysis more likely to be treated as qualified reviewers, editors, or models for subsequent work.

As contaminated publications accumulate, a positive feedback loop unfolds. Flawed syntheses are published in prestigious journals, cited as authoritative, used as models for subsequent reviews, and legitimated through peer review, potentially by scholars whose own work reflects the same errors. In an atmosphere of intense demand for evidence that AI works, popularity can be mistaken for validity, and repetition can be mistaken for consensus. Flawed meta-analyses do not just misstate an evidence base; they manufacture one.

5.6 Policy and societal implications

Meta-analyses are highly influential. Evidence hierarchies commonly place meta-analyses at the top, above randomized controlled trials (OCEBM Levels of Evidence Working Group, 2011; Wallace et al., 2022). Authors of AIED meta-analyses are aware of this prestige. Doo and Park (2026) write, “A meta-analysis enables researchers to synthesize individual quantitative research findings and provide more

robust and convincing conclusions” (p. 267). Wu et al. (2026) describe their own meta-analysis as a “scientific basis for policymaking” (p. 2). Because they have perceived prestige, flawed meta-analyses can be especially harmful (Ioannidis, 2016).

The social power of the AIED meta-analyses audited in this study also comes from the nature of the questions they claim to answer: “Does AI improve learning?” is nonsensical in educational terms but highly useful in market terms. It asks whether a product category works, not whether a pedagogical design works. AIED is in a policy environment where adoption is already treated as necessary and innovative, even as scientific evidence is lacking (Ro, 2025; U.S. Department of Education, 2026). As a result, there is a strong drive to produce evidence affirming the benefits of AI for education (O’Neill, in review). The audited meta-analyses are not shy about concluding that they have produced such evidence: “The comprehensive application of ChatGPT across all educational levels, subjects, and processes holds promise for improving student abilities across multiple dimensions” (Wu et al., 2026, p. 11); “Continuous use of ChatGPT should be ensured to support student learning” (Wang & Fan, 2025, p. 17); “This study provides a theoretical foundation for the future development of national GenAI education policies” (Liu et al., 2025, p. 14).

These recommendations are unwarranted given the poor methodology of the audited AIED meta-analyses. Worse, they are risky given the potential for AI to cause educational harm (Bastani et al., 2025; Georgiou, 2025; Kasneci et al., 2023; Lo et al., 2024). They can also misinform governments as they implement AI policies, such as the U.S. Department of Education’s 2026 final priority making AI literacy, educator AI training, and AI integration eligible for favorable treatment across discretionary grant competitions (U.S. Department of Education, 2026).

Flawed meta-analyses can transform weak and incoherent research into a policy-ready claim by aggregating bias and calling the result evidence. This is a particularly high risk in education because nearly all interventions show a positive result (Higgins et al., 2012). If a meta-analysis ignores

publication-bias adjustments, it removes the mechanisms that distinguish a real educational effect from accumulated research distortion. At that point, a pooled positive effect for “AI” is not a discovery that AI works. It reflects the ordinary positivity of education research due to novelty effects, publication bias, and small-study effects. In this way, flawed AIED meta-analyses launder bias into an apparently authoritative evidence base for AI adoption.

5.7 Peer review and editorial screening

Across the examined meta-analyses, broad claims about AIED suffered elementary validity failures that should have been detected before publication. Their recurrence shifts the concern from individual author error to the reliability of peer review, editorial screening, and journal standards for AIED meta-analysis. Because published meta-analyses can quickly become authoritative evidence for policy and practice, stronger safeguards are needed before such claims enter the literature.

6. Recommendations

6.1 Require full data transparency for meta-analysis

Authors should be required to share enough information to reproduce their results. For a meta-analysis this includes at least: search protocols, study selection decisions, coded study characteristics, extracted statistics, effect-size calculations, and analytic code if used. Authors should verify that all data needed to reproduce the pooled estimate, moderator analyses, heterogeneity statistics, and publication-bias tests are available. If a spreadsheet was used for data or calculations, it should be made available; if meta-analysis software was used the input data, settings, and output should be recorded each time it is used.

6.2 Peer review

Given the widespread methodological failure of published meta-analyses, editors should not consider a reviewer qualified based on having published a meta-analysis, even within a prestigious journal. An editor with statistical expertise should seek reviewers who have written skillful meta-

analyses; journals that lack editorial capacity to evaluate meta-analytic expertise should not publish meta-analyses without external statistical review. Peer reviewers must be competent to evaluate effect-size extraction, dependence, heterogeneity, publication bias, and moderator models. At least one peer reviewer should be widely read in the subject matter and asked to evaluate a subset of primary studies. In the present review, familiarity with the work of Mireille Bétrancourt was an asset in recognizing fabricated (likely AI-hallucinated) references attributed to her. Journals may wish to adopt a formal review instrument for peer reviewers to complete for meta-analysis.

6.3 Best practices when retraction occurs

Retractions are a necessary but minimally effective method of mitigating the damage done when a faulty study is published. Wang et al. (2022) found that retracted articles received most of their citations after retraction, and only 5% of articles citing them acknowledged the retraction. Retractions should follow the guidelines from the Committee on Publication Ethics (COPE). Retractions should be visible at every point where the article is discovered, accessed, exported, or cited. PDFs of the article should be watermarked “Retracted.” Citation exports and search results should include “[Retracted]” in the title field. Indexes like Crossref, PubMed, ERIC, Scopus, Web of Science and Google Scholar should be updated promptly. Articles that have cited the retracted article should have a note appended that the reference was retracted.

The retraction of Wang and Fan (2025) is a failure of both pre-publication review and of post-publication transparency. The letter requesting retraction of Wang and Fan (2025) was sent a month after its publication, yet the study circulated 351 days and accumulated more than 500 citations before being retracted. No expression of concern was published while the review was occurring, as advised by COPE if a resolution is expected to take “several months” or longer (COPE Council, 2025a, p. 3). This is necessary “To minimize harmful effects and uptake (e.g., citation of erroneous work, acting on their findings, or drawing incorrect conclusions)...” (COPE Council, 2025b, p. 11).

COPE also advises that a retraction note should state “the reasons and basis for retraction so that readers understand why the article is unreliable” (COPE Council, 2025b, p. 8). The retraction note is 45 words including that “The Editor has decided to retract the paper owing to concerns regarding discrepancies in the meta-analysis” (Editor, 2026, p. 1). This level of disclosure is insufficient: it prevents authors and reviewers from learning from the failures that occurred and it makes it impossible to fully correct the citation record because it may be unclear which uses of the article are implicated.

7. Limitations

An important limitation of the current study is that it focused on threshold-level validity failures, asking whether meta-analyses met minimum evidentiary standards. Deeper educational validity failures are also an important concern but were minimally assessed here (for such an assessment see: Bardach et al., 2025; Lawson et al., 2026). Even a meta-analysis that passes on threshold level validity might create an effect size that is not educationally meaningful, durable, or scalable. Experimental studies in education are prone to poor ecological validity, dependence on short-term measures, and novelty effects (Bronfenbrenner, 1976; Deaton & Cartwright, 2018), potentially confounding meta-analyses.

Only a portion of the primary studies selected for audit in this study were selected randomly. An overall prevalence of studies that created validity concerns was reported using only the randomly selected studies ($n = 28$, 61%). All other counts of primary studies include ones targeted for vetting and cannot be used as prevalence estimators.

This study did not attempt to identify every AIED meta-analysis published during the review period. The corpus was selected to examine a reproducible sample of recent meta-analyses making broad claims about AI, GenAI, or ChatGPT improving educational outcomes. The findings do not give the prevalence of validity failures across all AIED meta-analyses. However, the consistency of the findings across the examined corpus is itself consequential: all 14 reviewed meta-analyses contained severe validity failures, indicating that the problem is not isolated.

8. Conclusions

This study examined whether recent AIED meta-analyses provide valid support for broad claims that AI, GenAI, or ChatGPT improve educational outcomes. The answer is no.

All 14 audited meta-analyses suffered multiple, critical methodological failures. Constructs were incoherent, producing extreme heterogeneity that was not resolved. Violations of statistical independence were common, undermining pooled estimates, moderator analyses, heterogeneity analyses, and publication-bias tests. Moderator analyses that would be questionable with subgroups of $n = 10$ are run with subgroups of $n = 1$ and confidently converted into prescriptions for educational practice. Publication-bias assessment was inadequate: funnel plots were treated as meaningful evidence, low-powered tests were overinterpreted, critical values for Egger's test were not followed, dependence among effect sizes was often ignored, and discredited fail-safe N metrics were used as evidence of stability. No audited meta-analysis made a transparent and meaningful adjustment for publication bias, despite substantial reason to expect above-average publication bias in education technology (O'Neill, in review; Ro, 2025; Ropovik et al., 2021; Tamim et al., 2021). All these failures should have been identified during competent peer review, yet these articles were published in high-impact, peer-reviewed journals.

Publication bias is a fundamental threat to the validity of meta-analyses because it skews which studies are available. A competent meta-analysis must apply a robust method of adjustment for publication bias such as Bayesian approaches or selection models (Ropovik et al., 2021). Bartoš et al.'s (2025) critique of Wang and Fan (2025) showed the large effect size they reported entirely evaporated under multiple models of publication bias adjustment. This exemplifies how superficial dismissals of publication bias allow AIED meta-analyses to manufacture evidence by aggregating publication bias.

The audited meta-analyses are widely circulated. The high impact journals in which they appear are published by major academic presses. They have accumulated citations at an extraordinary rate.

Their conclusions circulate widely through subsequent literature as evidence that AI, GenAI, or ChatGPT improve learning. The problem is intensified by the current policy environment. AI adoption in education is being promoted as urgent, innovative, and increasingly inevitable, while schools and policymakers are being asked to act before a stable evidence base exists. In that context, these meta-analyses supply evidence that can be cited in grant proposals, policy briefs, product narratives, and implementation proposals.

The implications for peer review are severe. The failures documented in this audit are not based on theoretical disagreements; they are concrete errors. Some are easily visible within abstracts alone. Meta-analyses with incoherent constructs, violations of statistical assumptions of independence, underpowered and misinterpreted publication-bias tests, and drawing policy-facing conclusions from moderator analysis with subgroup sizes of one or two passed editorial and peer-review at prestigious journals. The gatekeeping systems that are supposed to protect the credibility of academic journals have failed.

References

- Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology*, 15(4), ep448. <https://doi.org/10.30935/cedtech/13417>
- Alneyadi, S., & Wardat, Y. (2024). Integrating ChatGPT in grade 12 quantum theory education: An exploratory study at Emirate school (UAE). *International Journal of Information and Education Technology*, 14(3), 398–410. <https://doi.org/10.18178/ijiet.2024.14.3.2061>
- Bachiri, Y.-A., Mouncif, H., & Bouikhalene, B. (2023). Artificial intelligence empowers gamification: Optimizing student engagement and learning outcomes in E-learning and MOOCs. *International Journal of Engineering Pedagogy (IJEP)*, 13(8), 4–19. <https://doi.org/10.3991/ijep.v13i8.40853>
- Bardach, L., Emslander, V., Kasneci, E., Eitel, A., Lindner, M., & Bailey, D. H. (2025). *Research Syntheses on AI in Education Offer Limited Educational Insights* (Dx6kt_v2). OSF Preprints. https://doi.org/10.31219/osf.io/dx6kt_v2
- Bartoš, F., Martinková, P., & Wagenmakers, E. (2025, May 12). Adjusting for publication bias reveals no evidence for the effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking. https://doi.org/10.31234/osf.io/8vs32_v1
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Becker, B. J., Hedges, L. V., & Pigott, T. D. (2004). *Campbell Collaboration statistical analysis policy brief*. Campbell Collaboration. http://www.campbellcollaboration.org/ECG/policy_statasp

- Beidel, D. C., Tuerk, P. W., Spitalnick, J., Bowers, C. A., & Morrison, K. (2021). Treating childhood social anxiety disorder with virtual environments and serious games: A randomized trial. *Behavior Therapy*, 52(6), 1351–1363. <https://doi.org/10.1016/j.beth.2021.03.003>
- Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1), 5–18. <https://doi.org/10.1002/jrsm.1230>
- Boyes, B. (2026, May 22). Disgraceful AI acts are growing. Will the knowledge management community be next? *RealKM*. <https://realkm.com/2026/05/22/disgraceful-ai-acts-are-growing-will-the-knowledge-management-community-be-next/>
- Bronfenbrenner, U. (1976, April). *The experimental ecology of education* [AERA Award Address]. American Educational Research Association Annual Meeting. ERIC. <https://eric.ed.gov/?id=ED131025>
- Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 5(8), 980–989. <https://doi.org/10.1038/s41562-021-01143-3>
- Conaway, C., Tipton, E., & Artiles, A. J. (2024). The Value of Variation: Embracing Heterogeneity in Intervention Research. *Teachers College Record*, 126(8), 107–116. <https://doi.org/10.1177/01614681241276956>
- Copas, J., & Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1(3), 247–262. <https://doi.org/10.1093/biostatistics/1.3.247>
- COPE Council. (2025a). Expressions of concern. Committee on Publication Ethics. <https://publicationethics.org/guidance/guideline/expressions-concern>
- COPE Council. (2025b). Retraction guidelines. Committee on Publication Ethics. <https://publicationethics.org/guidance/guideline/retraction-guidelines>

- Cuijpers, P., Griffin, J. W., & Furukawa, T. A. (2021). The lack of statistical power of subgroup analyses in meta-analyses: A cautionary note. *Epidemiology and Psychiatric Sciences*, 30, e78.
<https://doi.org/10.1017/S2045796021000664>
- Dai, C.-P., Ke, F., Pan, Y., Moon, J., & Liu, Z. (2024). Effects of artificial intelligence-powered virtual agents on learning outcomes in computer-based simulations: A meta-analysis. *Educational Psychology Review*, 36(1), Article 31. <https://doi.org/10.1007/s10648-024-09855-4>
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Dekkers, O. M. (2018). Meta-analysis: Key features, potentials and misunderstandings. *Research and Practice in Thrombosis and Haemostasis*, 2(4), 658–663. <https://doi.org/10.1002/rth2.12153>
- Doo, M. Y., & Park, Y. (2026). A meta-analysis of ChatGPT's influence on learning achievement. *International Review of Research in Open and Distributed Learning*, 27(1), 265–289.
<https://doi.org/10.19173/irrodl.v27i1.8775>
- Editor. (2026). Retraction note: the effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 13(1), 528. <https://doi.org/10.1057/s41599-026-07310-z>
- Fanelli, D. (2010). “Positive” results increase down the Hierarchy of the Sciences. *PLoS ONE*, 5(4), Article e10068. <https://doi.org/10.1371/journal.pone.0010068>
- Georgiou, G. P. (2025). *ChatGPT produces more “lazy” thinkers: Evidence of cognitive engagement decline* (arXiv:2507.00181). arXiv. <https://doi.org/10.48550/arXiv.2507.00181>
- Gu, J., & Yan, Z. (2025). Effects of GenAI interventions on student academic performance: A meta-analysis. *Journal of Educational Computing Research*, 63(6), 1460–1492.
<https://doi.org/10.1177/07356331251349620>

- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1), 39–65.
<https://doi.org/10.1002/jrsm.5>
- Heene, M. (2010). *A brief history of the fail safe number in applied research* (arXiv:1010.2326). arXiv.
<https://doi.org/10.48550/arXiv.1010.2326>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley-Blackwell.
- Higgins, S., Xiao, Z., & Katsipataki, M. (2012). *The impact of digital technology on learning: A summary for the Education Endowment Foundation. Full report*. Education Endowment Foundation.
<https://eric.ed.gov/?id=ED612174>
- Ingebrigtsen, M., & Lukic, M. (2025). *Substantial errors invalidate the conclusion from Wang and Fan (2025) about ChatGPT and learning outcomes*. PsyArXiv.
https://doi.org/10.31234/osf.io/wgu56_v1
- IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, 6(7), e010247.
<https://doi.org/10.1136/bmjopen-2015-010247>
- Ioannidis, J. P. A. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3), 485–514.
<https://doi.org/10.1111/1468-0009.12210>
- Ji, Y., Zhou, X., Li, T., & Zhan, Z. (2023). The effectiveness of ChatGPT on pre-service teachers' STEM teaching literacy, learning performance, and cognitive load in a teacher training course. *Proceedings of the 2023 6th International Conference on Educational Technology Management, ICETM '23*, 16–22. <https://doi.org/10.1145/3637907.3637948>

- Jin, Z.-C., Wu, C., Zhou, X.-H., & He, J. (2014). A modified regression method to test publication bias in meta-analyses with binary outcomes. *BMC Medical Research Methodology*, 14(1), 132.
<https://doi.org/10.1186/1471-2288-14-132>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1), 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial Intelligence–Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial. *JMIR Formative Research*, 5(8), e20678. <https://doi.org/10.2196/20678>
- Lawson, A. P., Martella, A. M., Weidlich, J., Mulders, M., & Buchner, J. (2026). Color Me Confounded: A Critical Analysis of Media Comparisons on ChatGPT in Education. *Computers & Education*, 105701. <https://doi.org/10.1016/j.compedu.2026.105701>
- Lee, D., & Kwon, H. (2024). Meta-analysis on effects of artificial intelligence education in K-12 South Korean classrooms. *Education and Information Technologies*, 29(17), 22859–22894.
<https://doi.org/10.1007/s10639-024-12738-4>
- Li, S., Zeng, C., Liu, H., Jia, J., Liang, M., Cha, Y., Lim, C. P., & Wu, X. (2025). A meta-analysis of AI-enabled personalized STEM education in schools. *International Journal of STEM Education*, 12(1), 58.
<https://doi.org/10.1186/s40594-025-00566-y>

- Lin, L., Chu, H., Murad, M. H., Hong, C., Qu, Z., Cole, S. R., & Chen, Y. (2018). Empirical comparison of publication-bias tests in meta-analysis. *Journal of General Internal Medicine*, 33(8), 1260–1267.
<https://doi.org/10.1007/s11606-018-4425-7>
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE Publications.
- Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions*, 27, Article 100495. <https://doi.org/10.1016/j.invent.2022.100495>
- Liu, Z., Zuo, H., & Lu, Y. (2025). The impact of ChatGPT on students' academic achievement: A meta-analysis. *Journal of Computer Assisted Learning*, 41(4), Article e70096.
<https://doi.org/10.1111/jcal.70096>
- Lo, C. K., Hew, K. F., & Jong, M. S. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, 219, 105100.
<https://doi.org/10.1016/j.compedu.2024.105100>
- Ma, N., & Zhong, Z. (2025). A meta-analysis of the impact of generative artificial intelligence on learning outcomes. *Journal of Computer Assisted Learning*, 41(5), Article e70117.
<https://doi.org/10.1111/jcal.70117>
- National Academies of Sciences, Engineering, and Medicine. (2018). How people learn II: Learners, contexts, and cultures. The National Academies Press. <https://doi.org/10.17226/24783>
- OCEBM Levels of Evidence Working Group. (2011). *The Oxford levels of evidence 2*. Oxford Centre for Evidence-Based Medicine. <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebmllevels-of-evidence>
- O'Neill, P. E. (in review). Framing artificial intelligence in education: Innovation narratives, contested equity, and environmental silence in journal calls for papers (CFPs). Manuscript submitted for publication.

Patience, N. (2026, Feb 12). AI capex 2026: The \$690B infrastructure sprint. *Futurum*.

<https://futurumgroup.com/insights/ai-capex-2026-the-690b-infrastructure-sprint/>

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, 295(6), 676–680.

<https://doi.org/10.1001/jama.295.6.676>

Reich, J. (2025, Oct 3). What past education technology failures can teach us about the future of AI in schools. *The Conversation*. <https://doi.org/10.64628/AAI.vvdntn96m>

Ro, J. (2025). Enforcing unwarranted optimism: Critical frame analysis on educational digitalisation policies in South Korea. *Learning, Media and Technology*, 1–16.

<https://doi.org/10.1080/17439884.2025.2462940>

Ropovik, I., Adamkovic, M., & Greger, D. (2021). Neglect of publication bias compromises meta-analyses of educational research. *PLoS ONE*, 16(6), e0252415.

<https://doi.org/10.1371/journal.pone.0252415>

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>

Sadek, M. R., & Mohamed, N. A. G. (2023). Artificial intelligence as a pedagogical tool for architectural education: What does the empirical evidence tell us? *MSA Engineering Journal*.

<https://doi.org/10.21608/msaeng.2023.291867>

Scargle, J. (2000). Publication bias: The “file-drawer” problem in scientific inference. *Journal of Scientific Exploration*, 14, 91–106.

Sinha, M. K., & Montori, V. M. (2006). Reporting bias and other biases affecting systematic reviews and meta-analyses: A methodological commentary. *Expert Review of Pharmacoeconomics & Outcomes Research*, 6(5), 603–611. <https://doi.org/10.1586/14737167.6.5.603>

- 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*, 343:d4002. <https://doi.org/10.1136/bmj.d4002>
- Sun, L., & Zhou, L. (2024). Does generative artificial intelligence improve the academic achievement of college students? A meta-analysis. *Journal of Educational Computing Research*, 62(7), 1896–1933. <https://doi.org/10.1177/07356331241277937>
- Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *BMJ : British Medical Journal*, 320(7249), 1574–1577. <https://doi.org/10.1136/bmj.320.7249.1574>
- Tamim, R. M., Borokhovski, E., Bernard, R. M., Schmid, R. F., Abrami, P. C., & Pickup, D. I. (2021). A study of meta-analyses reporting quality in the large and expanding literature of educational technology. *Australasian Journal of Educational Technology*, 37(4), 100–115. <https://doi.org/10.14742/ajet.6322>
- Tipton, E., Bryan, C., Murray, J., McDaniel, M. A., Schneider, B., & Yeager, D. S. (2023). Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity: Commentary on Macnamara and Burgoyne (2023) and Burnette et al. (2023). *Psychological Bulletin*, 149(3–4), 229–241. <https://doi.org/10.1037/bul0000384>
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, 10(2), 161–179. <https://doi.org/10.1002/jrsm.1338>

- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, 10(2), 180–194.
<https://doi.org/10.1002/jrsm.1339>
- Tuomi, I. (2025). What counts as evidence in AI & ED: Towards science-for-policy 3.0. *European Journal of Education Policy and Practice*, 1(1), 1–31. <https://doi.org/10.5117/EJEP2025.1.001.TUOMI>
- U.S. Department of Education. (2026). *Final priority and definitions—Secretary’s supplemental priority and definitions on advancing artificial intelligence in education*. *Federal Register*, 91, 18774–18780.
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10(4), 428–443. <https://doi.org/10.1037/1082-989X.10.4.428>
- Wallace, S. S., Barak, G., Truong, G., & Parker, M. W. (2022). *Hierarchy of evidence within the medical literature*. *Hospital Pediatrics*, 12(8), 745–750. <https://doi.org/10.1542/hpeds.2022-006690>
- Wallace-Wells, D. (2026, May 20). What schools are forgetting in their race to embrace A.I. *The New York Times*. <https://www.nytimes.com/2026/05/20/opinion/ai-schools-backlash.html>
- Wang, J., & Fan, W. (2025). The effect of ChatGPT on students’ learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis [Retracted]. *Humanities and Social Sciences Communications*, 12(1), Article 621. <https://doi.org/10.1057/s41599-025-04787-y>
- Wang, Z., Shi, Q., Zhou, Q., Zhao, S., Hou, R., Lu, S., Gao, X., & Chen, Y. (2022). Retracted systematic reviews continued to be frequently cited: A citation analysis. *Journal of Clinical Epidemiology*, 149, 137–145. <https://doi.org/10.1016/j.jclinepi.2022.05.013>

- Weidlich, J., Gašević, D., Drachsler, H., & Kirschner, P. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning*, 41(5), Article e70105.
<https://doi.org/10.1111/jcal.70105>
- Winter, J. (2026, April 23). What will it take to get A.I. out of schools? *The New Yorker*.
<https://www.newyorker.com/culture/progress-report/what-will-it-take-to-get-ai-out-of-schools>
- 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>
- Wu, W., Duan, J., Reed, W. R., & Tipton, E. (2025). What can we learn from 1,000 meta-analyses across 10 different disciplines? *Research Synthesis Methods*, 17(1), 123–156.
<https://doi.org/10.1017/rsm.2025.10035>
- Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications*, 13(1), 684. <https://doi.org/10.1057/s41599-026-07019-z>
- Yang, T.-C., Hsu, Y.-C., & Wu, J.-Y. (2025). The effectiveness of ChatGPT in assisting high school students in programming learning: Evidence from a quasi-experimental research. *Interactive Learning Environments*, 33(6), 3726–3743. <https://doi.org/10.1080/10494820.2025.2450659>
- Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence*, 4, 100147.
<https://doi.org/10.1016/j.caeai.2023.100147>
- Zhao, Y., Yue, Y., Sun, Z., Jiang, Q., & Li, G. (2025). Does generative artificial intelligence improve students' higher-order thinking? A meta-analysis based on 29 experiments and quasi-experiments. *Journal of Intelligence*, 13(12), 160. <https://doi.org/10.3390/jintelligence13120160>

- Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2023). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. *Interactive Learning Environments*, 31(9), 5650–5664. <https://doi.org/10.1080/10494820.2021.2015693>
- Zhu, Y., Liu, Q., & Zhao, L. (2025). Exploring the impact of generative artificial intelligence on students' learning outcomes: A meta-analysis. *Education and Information Technologies*, 30(11), 16211–16239. <https://doi.org/10.1007/s10639-025-13420-z>