Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Independent Learning after Generative AI Support in School Mathematics: A Systematic Review and Assessment-Stratified Meta-Analysis

Alexander Kuznetsov

alex@5steps.edu.sg

Center for Educational & Psychological Assessment, 5 Steps Academy Private Limited
https://orcid.org/0009-0002-6652-3544

Systematic Review

Keywords: generative artificial intelligence, large language models, K–12 mathematics education, assessment conditions, cognitive offloading, productive friction, systematic review, meta-analysis

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11766988/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Independent Learning after Generative AI Support in School Mathematics: A Systematic Review and Assessment-Stratified Meta-Analysis

Alexander Kuznetsov, Ph.D.
Center for Educational & Psychological Assessment, 5 Steps Academy Private Limited, Singapore | alex@5steps.edu.sg | ORCID: 0009-0002-6652-3544

Abstract

Background and Objective. Generative artificial intelligence (GenAI) solves school math problems on demand. Higher scores during live tool access do not prove retained mathematical competence once that support disappears. Earlier meta-analyses pooled assisted task completion together with unmonitored and closed-book posttests, which obscured whether pupils actually learned to work alone. This PRISMA 2020 systematic review and stratified meta-analysis evaluates K–12 mathematics outcomes by separating test-time resource rules from instructional design.

Methods. Two-stream searches across ERIC, APA PsycNet, OpenAlex, Crossref, Semantic Scholar, arXiv, EuropePMC, and bidirectional citation chasing (1 January 2022–8 November 2026) returned 11,061 raw entries and 6,530 unique records. Every sought report was retrieved in full (372 / 372; 0 unretrieved). Eligibility verification and mathematical consistency checks retained 49 K–12 school-mathematics reports across 46 independent cohorts and supplied 36 quantitative contrasts across 20 controlled study families (21 controlled quantitative reports, plus 1 factorial within-AI modality trial synthesized narratively) alongside 28 qualitative, single-arm, observational, and boundary-comparator reports. Random-effects Paule–Mandel and restricted maximum likelihood estimators with classroom cluster adjustments (calibrated intraclass correlation $\rho = 0.128$) stratified contrasts across Category A (GenAI active during testing), Category B (posttest with undocumented resource rules), Category C (partial human-help restriction), and Category D (explicit closed-resource immediate and delayed exams).

Results. Within primary multi-cluster and individually randomized trials ($k = 14$), large concurrent gains during live GenAI access (Category A: $g = +0.615$, 95% CI [+0.044, +1.185], $p = .035$) shrank across undocumented posttests (Category B: $g = +0.186$ [+0.009, +0.362]) and tutor-restricted exit tickets (Category C: $g = +0.095$ [+0.024, +0.167]) down to a tight null on explicit closed-resource exams (Category D: $g = −0.010$ [−0.137, +0.117], $p = .876$; focal contrast $\Delta g = +0.625$, $p = .036$). Extended controlled trials ($k = 28$) confirmed strong between-category moderation (Q_between = 12.55, $p = .0059$; Category D: $g = −0.035$ [−0.149, +0.078]). Pedagogical architecture accounted for all primary post-withdrawal heterogeneity (Q_between = 18.69, $p = .0001$; $I^2 = 0\%$ within every tier). Unconstrained direct-answer access harmed closed-exam scores (M1: $g = −0.205$, $p = .043$), passive hint guardrails blocked harm without yielding net transfer (M2: $g = −0.038$, $p = .173$), and architectures enforcing productive cognitive friction (unaided mastery gating, teachable agents, Socratic Pólya scaffolding, and human-in-the-loop oversight) produced durable gains (M3: $g = +0.100$, $p = .0001$; extended $g = +0.315$, $p = .0002$). Head-to-head within-AI trials ($k = 6$) favored pedagogical guardrails and teacher mediation over unguarded GenAI ($g = +0.645$ [+0.215, +1.075], $p = .0033$), whereas upper-secondary algebra and calculus cohorts showed acute offloading and working-memory overload.

Conclusions. Unstratified, unclustered synthesis inflates verified closed-resource mathematics learning by +0.26 to +0.29 SD (naive $g = +0.275$ vs. closed-resource $g = −0.010$). Durable competence requires crossing two distinct thresholds. Instructional systems must block direct answer copying and enforce active learner generation before and after AI withdrawal.

Keywords: generative artificial intelligence; large language models; K–12 mathematics education; assessment conditions; cognitive offloading; productive friction; systematic review; meta-analysis

1. Introduction

1.1 Assessment Conditions Define the Estimand: Effects with AI Versus Effects of AI

Salomon et al. (1991) separated *effects with technology* (joint performance while a learner and an external tool work together) from *effects of technology* (the cognitive capability that remains when the learner later works alone). Conversational large language models (LLMs) and generative artificial intelligence (GenAI) sharpen that classic boundary in K–12 mathematics. Classical step-based intelligent tutoring evaluations routinely tested post-instructional mastery on unassisted exams (Ma et al., 2014), whereas conversational chatbots can supply complete symbolic solutions during practice and testing alike.

Blurring those two estimands corrupts causal inference. Aggregate averages mislead. Early syntheses of ChatGPT and educational GenAI (Boolzen et al., 2026; Deng et al., 2025) reported moderate-to-large achievement gains, yet those averages combined supported task completion during live chatbot access with classroom posttests lacking documented invigilation rules and omitted variance adjustments for intact-classroom assignment.

1.2 Cognitive Offloading, the Fluency Illusion, and Enforced Productive Friction

Three cognitive mechanisms explain why unmediated GenAI access often fails to build, and sometimes erodes, independent mathematical skill. Cognitive offloading theory (Risko & Gilbert, 2016) holds that learners delegate effortful mental steps to reliable external tools whenever immediate task completion carries the primary reward. In school mathematics classrooms, external delegation of equation setup, algebraic manipulation, or derivative rules short-circuits the retrieval and schema assembly needed for durable memory (Sweller et al., 2019). Even attentive reading of a coherent LLM walkthrough can trigger a fluency illusion (Bjork & Bjork, 2011). Smooth step-by-step prose lowers perceived effort during study, so pupils mistake comprehension of someone else's derivation for personal mastery and curtail self-testing (Lademann et al., 2025). Research on desirable difficulties and productive failure (Bjork & Bjork, 2011; Kapur, 2016) points to the missing ingredient: durable transfer requires students to construct representations, wrestle with impasses, diagnose errors, and solve fresh problems unassisted.

A non-linear pedagogical prediction follows directly from these three mechanisms. Simple contrasts between 'unguarded chatbots' and 'prompted AI tutors' miss the mark. Telling an LLM via a system prompt to withhold final numerical answers and dispense hints should stop blatant copy-and-paste offloading and prevent negative exam harm relative to unassisted study. Passive hint consumption still demands very little generative effort. Unless the surrounding workflow requires pupils to complete consecutive unaided mastery streaks, teach a fallible agent, critique deliberate errors, or justify steps to a classroom teacher, net post-withdrawal gains over standard instruction should stay near zero.

1.3 Research Questions

We test these predictions across the global empirical record in K–12 school mathematics (2022–2026). Four research questions organize the synthesis:

- **RQ1 (Assessment-Condition Attenuation):** How do pooled standardized effects of GenAI mathematics instruction shift across test-time resource regimes, from live supported performance (Category A) through undocumented and partially restricted posttests (Categories B and C) to explicit closed-resource immediate and delayed examinations (Category D)?

- **RQ2 (Pedagogical Architecture and Within-AI Contrasts):** Across post-withdrawal assessments, how do unconstrained direct-answer access (M1), passive hint guardrails or static explanations (M2), and enforced productive cognitive friction (M3) moderate transfer, and what do head-to-head within-AI experiments and expert teacher benchmarks show?
- **RQ3 (Developmental Stage and Classroom Process):** How does post-withdrawal transfer vary across primary (Grades 3–5), middle (Grades 6–8), and senior high school (Grades 9–12/13) cohorts, and what interactional dynamics surface in qualitative discourse and observational classroom studies?
- **RQ4 (Methodological Counterfactual):** By what margin does conventional unstratified, unclustered meta-analysis misstate the effect of GenAI on verified independent mathematics learning?

2. Methods

2.1 Protocol Scope, Two-Stream Search Architecture, and Complete Full-Text Retrieval

Reporting follows PRISMA 2020 standards (Page et al., 2021). No prospective registration is claimed. This full-retrieval synthesis united a core bibliographic register with programmatic multi-index discovery spanning 1 January 2022 through 8 November 2026.

Two fully documented search streams supplied candidate records (verbatim query syntax appears in Supplement S1, §S1.1). Stream 1 combined eight structured ERIC queries (E01–E08, executed 8 October 2026; 451 raw hits, 299 unique ERIC identifiers), a Boolean search in APA PsycNet (Any Field, executed 14 October 2026; 759 raw hits, 758 unique records), and nine early scoping seed papers that totaled 1,219 raw entries and 1,039 deduplicated core records (272 ERIC-only, 731 APA-only, 27 shared across both databases, and 9 seed-only). Stream 2 broadened the net. Executed on 8 November 2026 across six scholarly APIs (OpenAlex, Crossref, Semantic Scholar, the full-window ERIC API, arXiv, and EuropePMC) alongside forward and backward citation chasing from anchor trials and prior reviews (Bastani et al., 2025; Boolzen et al., 2026; Deng et al., 2025; Z. Liu et al., 2025; R. E. Wang et al., 2025). Removing 4,351 internal and cross-stream overlaps from the 9,842 raw Stream-2 hits (4,531 duplicates overall, including 180 within Stream 1) left 5,491 net new records and a combined screening pool of 6,530 unique records.

Every sought full text was retrieved. Across all 372 candidate reports slated for full-text inspection, unretrieved reports stood at zero (100.0% retrieval). All 30 scoping candidates from Stream 1 and all 14 subscription-access K–12 mathematics articles from Stream 2 were acquired via publisher subscriptions, institutional archives, or verified author project repositories (item-by-item ledgers appear in Supplement S1, §S1.2).

2.2 Eligibility Criteria (PICOS), Cohort Linkage, and Forensic Numerical Gate

Four PICOS and integrity rules governed study inclusion:

- **Population:** Enrolled primary, middle, or secondary school pupils (Grades K–12, plus Grade 13 / sixth-form upper-secondary equivalents) formed the target population. University, adult crowdworker, and preschool samples were excluded.
- **Intervention / Exposure:** Eligible studies examined instructional, tutoring, scaffolding, feedback, problem-posing, or self-directed mathematics learning with a generative AI model (ChatGPT, GPT-3.5/4/4o/5, Claude, Gemini, Llama, Qwen, or LLM-driven teachable and tutoring agents). Deterministic rule-based tutoring platforms without generative language or multimodal models were excluded.
- **Outcomes and Design:** Eligible designs reported empirical pupil mathematics achievement, problem-solving accuracy, mathematical discourse or modeling processes, or directly linked learning dispositions across randomized controlled trials (RCTs), controlled quasi-experiments, longitudinal natural experiments, single-group classroom interventions, cross-sectional surveys, or qualitative case

studies. Pure solver or grading benchmarks without human K–12 pupils, teacher-only professional development studies lacking student outcomes, inseparable general STEM or web-programming courses (such as HTML/CSS instruction in López-Pernas et al., 2026), and conceptual essays (Kamaghe, 2026) were excluded.

- **Forensic Numerical Integrity Gate (for quantitative meta-analysis):** Arithmetic consistency came first. Every candidate controlled study underwent script-based statcheck verification of means, standard deviations, sample sizes, degrees of freedom, and test statistics (t, F, Z, χ^2, odds ratios). Full-text auditing excluded reports with irreconcilable mathematical contradictions: one-way ANOVA F-ratios diverging more than twofold from printed group means and standard deviations alongside anachronistic references (Chen, 2025), confounded multi-platform bundles that swapped arm sample sizes between text and tables (Canonigo, 2024; Polydoros, 2025), and unverifiable synthetic variance structures (4 statcheck exclusions plus 1 duplicate dissertation = 5 in the PRISMA integrity/duplicate gate, alongside 1 uncontrolled multi-platform bundle placed in the n = 12 confounded-bundle category; Supplement S1, §S1.3). Companion papers drawn from identical learner cohorts were linked explicitly to prevent unit-of-analysis duplication. Journal and conference versions of ChatGPT-MPS reporting identical paired statistics on 49 elementary pupils (J. Liu et al., 2025a, 2025b) and their usability report (J. Liu et al., 2025c), as well as the ALTER-Math controlled study (Xing et al., 2025), its single-group log analysis (N = 589; Z. Liu et al., 2025), and its subsequent voice-personalization RCT (Li et al., 2026), were cross-indexed so no pupil entered any pooled model twice.

2.3 Four-Category Assessment-Condition Taxonomy and Pedagogical Architecture Tiers

Test rules govern construct meaning. Because a mathematics score reflects whatever assistance was permitted during administration, every quantitative contrast was assigned a priori to one of four mutually exclusive assessment-condition categories:

- **Category A: Supported Performance (GenAI Active During Assessment).** Accuracy, task completion, or unproctored mastery is recorded while pupils retain live access to the generative AI system or embedded LLM hints.
- **Category B: Immediate Posttest with Unresolved Resource Access.** A post-intervention mathematics test follows instruction, but the primary report documents neither proctoring nor device lock-down nor explicit rules barring AI and external notes during testing.
- **Category C: Posttest with Partial Human-Help or Platform-Hint Restriction.** Post-instructional testing enforces a documented ban on tutor assistance or disables in-platform AI hints (such as session exit tickets under a 'no-tutor-help' rule), while external student resources in remote or classroom settings remain unmonitored.
- **Category D: Explicit Closed-Resource Immediate and Delayed Assessments.** Invigilated paper-and-pencil exams, locked-device tests, or proctored retention checks bar GenAI, external devices, and unauthorized help. Independently of test-time rules, each intervention's study-phase pedagogy was coded into three tiers: **M1 (Direct-Answer Offloading)**, where pupils query an unconstrained LLM that outputs full solutions on request; **M2 (Passive Hint Guardrails or Static Explanations)**, where system prompts withhold final answers or display structured AI explanations without requiring unaided mastery; and **M3 (Enforced Productive Friction)**, where the workflow mandates active learner generation through consecutive unaided mastery streaks (3-in-a-row), fallible AI peer teaching, deliberate error debugging, Socratic Pólya step gating, or human-in-the-loop teacher oversight.

Design rigor varied sharply. What Works Clearinghouse (WWC) v5.0 criteria (What Works Clearinghouse, 2022) separated GenAI-versus-non-AI contrasts into **Tier-1 Primary studies** (multi-cluster trials with $K \geq 4$ classrooms or schools, or individual randomization free of internal arm contradictions) and **Tier-2 Sensitivity studies** (single-cluster-per-arm K = 2 classroom designs where treatment is algebraically confounded with

teacher identity, cohort comparisons with control-label ambiguity, unreviewed preprints, or topic-level difference-in-differences natural experiments). **Direct Within-AI Guardrail Contrasts** (comparing two GenAI arms under different pedagogical rules) and **Active Expert Human Benchmarks** (comparing GenAI explanations or feedback against experienced mathematics teachers) were estimated in dedicated models rather than mixed with business-as-usual controls.

2.4 Effect-Size Derivation, Cluster Adjustment, and Meta-Analytic Estimation

Scale comparability required uniform standardization. Continuous outcomes were standardized as small-sample bias-corrected Hedges' g (Borenstein et al., 2021). When commensurate baseline and posttest means and standard deviations were available, Morris's pretest-adjusted g_ppc2 (Morris, 2008) accounted for baseline imbalance; when ANCOVA adjusted differences or cluster-robust regression slopes (β) were reported, g was scaled by the pooled raw standard deviation and verified directly against student-level CSV microdata in Bastani et al. (2025). Binary mastery and log-odds outcomes used the Cox transformation ($g = \ln(OR) \times \sqrt{3} / \pi$; Cox, 1970), while skewed Mann–Whitney rank outcomes in Luo et al. (2026) used exact area-under-the-curve conversion (AUC → $\sqrt{2}\,\Phi^{-1}$(AUC); Ruscio, 2008).

Classroom clustering changes inference. Whenever cluster-assigned classroom or school studies omitted cluster-robust standard errors, sampling variances were multiplied by the WWC/Cochrane design effect ($DEFF = 1 + (\bar{m} - 1)\rho$; Hedges, 2007; What Works Clearinghouse, 2022). Rather than assuming a conventional guess, we calibrated the primary classroom ICC (ρ = 0.128) via one-way random-effects ANOVA across the 44 non-honors classrooms in the open Part-3 closed-exam microdata of Bastani et al. (2025) and ran sensitivity checks at ρ = 0.05 and ρ = 0.20. In three-arm trials sharing a single control group (Bastani et al., 2025; Luo et al., 2026; Polydoros, 2026), control sample sizes were split (n_C / 2) in pooled syntheses to avoid double-counting reference learners.

Between-study variance τ^2 was estimated primarily via Paule–Mandel (PM; Paule & Mandel, 1982), paired with restricted maximum likelihood (REML) point checks, Hartung–Knapp–Sidik–Jonkman (HKSJ) t-intervals (Hartung & Knapp, 2001), and 95% prediction intervals. Subgroup moderation was evaluated with both separate-τ^2 Q_between tests and common-τ^2 mixed-effects Q_M omnibus tests, plus a pre-specified focal contrast between Supported Performance (Category A) and Closed-Resource Performance (Category D). Publication bias checks followed. Small-study asymmetry and outlier influence were checked using Egger's regression (Egger et al., 1997), Duval and Tweedie's L0 Trim-and-Fill (Duval & Tweedie, 2000), and Leave-One-Out (LOO) refitting via standalone Python (meta_stats_engine.py, run_final_meta_analysis.py) and R (reproduce_meta_analysis.R) scripts.

2.5 Critical Appraisal and Review Governance

Methodological quality was judged at the specific outcome-contrast level rather than via holistic paper scores. Individual RCTs were evaluated with Cochrane RoB 2 (Higgins et al., 2019), cluster RCTs with the RoB 2 cluster extension (Eldridge et al., 2021), and quasi-experimental, analytical cross-sectional (26 March 2024 manual), and qualitative studies with Joanna Briggs Institute checklists (JBI, 2024a, 2024b, 2025; domain matrices appear in Supplement S1, §S1.5). Scripted routines handled record deduplication and cross-checked extracted table figures against primary full texts.

3. Results

3.1 Study Selection, Full-Text Resolution, and Map of the 49 Included Reports

Figure 1 maps the PRISMA 2020 selection flow. Removing 4,531 duplicates from 11,061 raw entries across Streams 1 and 2 left 6,530 unique records for title and abstract screening. Screening excluded 5,940 off-topic or non-K–12 records, catalogued 218 secondary reviews, and advanced 372 candidate reports to full-text evaluation. All 372 reports were retrieved (0 unretrieved). Full-text review excluded 323 reports across seven

categories: post-secondary, adult, or preschool samples (n = 134); standalone LLM solver or grading benchmarks without human students (n = 79); teacher professional-development or lesson-design studies lacking pupil math outcomes (n = 48); non-mathematics or inseparable general STEM and coding courses (n = 34); deterministic rule-based tutors or confounded multi-platform bundles (n = 12); conceptual papers or protocols (n = 11); and statcheck integrity failures or duplicate dissertations (n = 5).

Figure 1. *PRISMA 2020 Flow Diagram of Two-Stream Identification, Complete Retrieval, Eligibility, and Study Inclusion*

PRISMA 2020 Phase	Main Identification, Screening, Retrieval & Inclusion Flow	Excluded / Unretrieved Records & Breakdown
1. Identification	**Stream 1 — Core Databases & Seed Register (Oct 2026):** ERIC E01–E08 (n = 451 raw; 299 unique), APA PsycNet (n = 759 raw; 758 unique), early scoping seed register (n = 9) → 1,219 raw records → **1,039 core unique records** (272 ERIC-only, 731 APA-only, 27 overlap, 9 seed). **Stream 2 — Multi-API & Citation Expansion (Nov 2026):** OpenAlex, Crossref, Semantic Scholar, ERIC API, arXiv & EuropePMC (2022–2026) + citation chasing (Bastani et al., 2025; Boolzen et al., 2026; Z. Liu et al., 2025; R. E. Wang et al., 2025) → 9,842 raw records → **5,491 additional unique records**. **Combined Deduplicated Total: 6,530 unique records** (from 11,061 raw records).	**Duplicates Removed Before Screening (n = 4,531):** Stream-1 internal duplicates (n = 180); Stream-2 & cross-stream overlaps (n = 4,351).
2. Screening	**Records Screened at Title & Abstract Level (n = 6,530):** Evaluated against K–12 school mathematics (Grades K–12 / 1–13) and generative AI / LLM eligibility criteria; all Stage-1 candidate records advanced to full primary-source audit.	**Excluded at Title / Abstract (n = 6,158):** Off-topic, higher-education, non-GenAI, or non-mathematics (n = 5,940); secondary reviews, meta-analyses & commentaries catalogued for citation chasing only (n = 218).
3. Retrieval	**Reports Sought for Full-Text Retrieval (n = 372):** 64 K–12 candidate reports from Core Registry + 308 from Multi-API & citation streams; **372 retrieved in full text** via open-access archives, institutional/publisher access, dissertation repositories & author deposits.	**Reports Not Retrieved (n = 0 — 100.0% Retrieval):** Zero unretrieved reports across all 372 sought full texts.
4. Eligibility	**Full-Text Reports Assessed for Eligibility & Statistical Integrity (n = 372):** Verified K–12 learner population, isolable mathematics outcome & generative AI component; classified under Assessment-Condition Taxonomy (Categories A–D) & Pedagogical Tiers (M1–M3); audited via statcheck recalculation & cohort-linkage gate.	**Full-Text Reports Excluded (n = 323):** Post-secondary/adult/preschool sample (n = 134); standalone LLM solver/grader benchmark without learners (n = 79); teacher PD/lesson-planning only (n = 48); non-mathematics or inseparable STEM/CS, e.g., López-Pernas et al., 2026 (n = 34); non-GenAI ITS or confounded bundle, e.g., Canonigo, 2024; Polydoros, 2025, Ch. 9 (n = 12); conceptual/protocol/commentary, e.g., Kamaghe, 2026 (n = 11); critical statcheck failure or duplicate dissertation, e.g., Chen, 2025; Polydoros et al., 2025; Nangolo, 2026 (n = 5).
5. Included	**Total Included Corpus (n = 49 Reports; 46 Independent Study Cohorts):** (1) **Quantitative Controlled Corpus — 21 Controlled Reports** (k =	**Cohort Deduplication Verified:** 49 included reports mapped to 46 independent learner cohorts; zero overlapping pupils double-

PRISMA 2020 Phase	Main Identification, Screening, Retrieval & Inclusion Flow	Excluded / Unretrieved Records & Breakdown
	36 extracted contrasts across 20 controlled study families + 1 narrative factorial voice-design RCT, Li et al., 2026): Tier-1 Primary GenAI vs. Non-AI (8 studies, k = 14: Cat A k = 3, Cat B k = 5, Cat C k = 1, Cat D k = 5); Tier-2 Sensitivity GenAI vs. Non-AI (9 studies, k = 14; Extended total: 17 studies, k = 28); Direct Within-AI Guardrails (k = 6 across 5 trials) & Active Expert Human Benchmarks (k = 2: Worden et al., 2025, Exp. 2; Cosentino et al., 2025). (2) **Qualitative, Single-Arm, Observational & Boundary Corpus — 28 Reports:** Discourse/commognitive case studies (n = 12); single-group pre–post / log-strata reports (n = 5, incl. linked J. Liu et al., 2025a, 2025b, 2025c, & Z. Liu et al., 2025); cross-sectional surveys (n = 2: Besas et al., 2026; Lintner et al., 2026); affective/rating studies (n = 3); static/offline boundary comparators (n = 6).	counted in any meta-analytic model.

Note. All 30 Stream-1 candidate records and all 14 Stream-2 subscription-access K–12 mathematics articles were retrieved and audited from complete primary texts (Reports not retrieved = 0).

Forty-nine reports covering 46 independent learner cohorts met inclusion criteria (Table 1). Across 19 countries in East and Southeast Asia (China, Taiwan, South Korea, Vietnam, Indonesia, Philippines), Europe (Germany, Norway, Greece, Czechia, Italy, Portugal), North America (United States), the Middle East (Turkey), Sub-Saharan Africa (Ghana), Central Asia (Kazakhstan), and South Asia (India), the evidence spans primary (Grades 3–6), middle (Grades 6–9), and upper-secondary (Grades 9–13) curricula. Controlled quantitative designs came from 20 study families across 21 reports: 20 reports supplied the 36 extracted contrasts (14 Tier-1 GenAI-vs.-non-AI contrasts across 8 studies, 14 Tier-2 GenAI-vs.-non-AI contrasts across 9 studies, 6 direct within-AI guardrail contrasts, and 2 active expert-teacher benchmarks), while 1 factorial within-AI voice-design RCT was synthesized narratively inside its ALTER-Math cohort family. Twenty-eight additional reports (26 cohorts) contributed qualitative discourse analyses, single-group pre–post studies, cross-sectional surveys, affective ratings, or static/offline boundary comparisons.

Table 1. *Complete Overview of the 49 Included School-Mathematics Reports Across Assessment Categories and Synthesis Roles (46 Independent Cohorts)*

Report	Sample & Setting	Design & Generative AI Modality	Cat.	Synthesis Role
Bastani et al. (2025)	Grades 9–11, Turkey (N = 825 pupils, 44 classes)	Cluster RCT (GPT-4 Base & GPT-4 Hint Tutor vs. textbook practice)	Cat A & Cat D	Tier-1 Meta (A-01a/b, D-01a/b) & Within-AI (W-01)
Oreopoulos et al. (2026)	Grades 6–8, USA (N = 6,327 across 2 strata)	2×2×2 Individual Login RCT (NUMI post-error AI ± mastery gating)	Cat A & Cat D	Tier-1 Meta (A-02, D-02a/b) & Within-AI (W-02)
Henkel et al. (2024)	Grades 3–9, Ghana (N = 478, 11 schools)	8-month School-Clustered Trial (Rori hybrid LLM WhatsApp tutor)	Cat D	Tier-1 Meta (D-03; paper exam)
R. E. Wang et al. (2025)	Grades 3–8, USA (900 tutors, 4,136 sessions)	Tutor-Clustered RCT (Tutor CoPilot real-time GPT-4 coaching)	Cat C	Tier-1 Meta (C-01; no-tutor-help exit ticket)
Kuo, Bai, & Lin (2025)	Grade 5, Taiwan (Exp 1 N = 363; Exp 2 N = 104)	Multi-school Quasi-Exp. ANCOVA (TALPer GPT-4o Socratic + Pólya)	Cat B	Tier-1 Meta (B-01a Exp 1, B-01b Exp 2)
Xing et al. (2025)	Middle School, USA (Study 3 N = 212, 4 classes)	5-week Quasi-Exp. Pre–Post Control (ALTER-Math teachable agent)	Cat B	Tier-1 Meta (B-02; ALTER-Math family)
Tasdelen & Bodemer (2025)	Grades 3–4, Germany (N = 114)	Individual Classroom RCT (GPT context-personalised word problems)	Cat B	Tier-1 Meta (B-03)
Lademann et al. (2025)	Grade 6, Germany (N = 214)	Individual RCT (custom ChatGPT explanations vs. textbook)	Cat B	Tier-1 Meta (B-04; fluency illusion)
Rismanchian et al. (2026)	Grades 5–12 & Gateway, USA (15.4M interactions)	Longitudinal Topic-FE DiD across 11 post-ChatGPT quarters (ALEKS)	Cat A & Cat D	Tier-2 Meta (A-03 unproctored, D-04 proctored)
H. Liu et al. (2026)	Grades 5, 8, 11, China (3 × N = 100 urban/rural)	3-Cohort 2×2 Quasi-Exp. + Eye-Tracking (Multimodal LLAMA-13B)	Cat B	Tier-2 Meta (B-05a G5, B-05b G8, B-05c G11)
Xuan et al. (2025)	Grade 11 (Limits), Vietnam (2 × N = 60)	Dual-School Pre–Post Control (Private FPT & Public Thoi Thuan)	Cat B	Tier-2 Meta (B-06a Private, B-06b Public)
Lin et al. (2025)	Grade 5, Taiwan (N = 59, 2 classes)	Intact-Class Quasi-Exp. (debugging deliberate GPT-4 errors)	Cat B	Tier-2 Meta (B-07; K = 2 classes)
Karaman & Göksu (2024)	Grade 3, Turkey (N = 39, 2 classes)	5-week Intact-Class Quasi-Exp. (ChatGPT lesson plans vs. teacher)	Cat B	Tier-2 Meta (B-08; K = 2 classes)
Malik et al. (2026)	Grade 8, Indonesia (N = 47, 2 classes)	Intact-Class Quasi-Exp. (Pólya-aligned ChatGPT prompting)	Cat B	Tier-2 Meta (B-09; K = 2 classes)
Davor (2026)	Senior High, Ghana (N = 86, 2 classes)	Intact-Class Quasi-Exp. (GenAI cognitive scaffold in geometry)	Cat B	Tier-2 Meta (B-10; baseline table swap)
Prabhakar (2026)	Grade 7, India (N = 100)	Intact-Classroom Quasi-Exp. (GenAI dialogic thinking partner)	Cat B	Tier-2 Meta (B-11; non-indexed venue)
Polydoros (2026)	Grade 7, Greece (N = 396, 9 schools)	3-Arm Quasi-Exp. Preprint (ChatGPT-only, Blended Teacher+AI, Typical)	Cat B	Tier-2 Meta (B-12a/b) & Within-AI (W-05)
Worden et al. (2025)	Grades 6–8, USA (Exp 1 N = 188; Exp 2 N = 6,371)	Within-Platform ASSISTments RCTs (GPT-4 CoT vs. Answer & vs. Teacher)	Cat C	Within-AI (W-03) & Human Benchmark (H-01)
Luo, Wang, & Hew (2026)	Grade 5, China (Study 2 N = 95 on Week-7 exam)	3-Arm Trial (Teacher HITL+SRL vs. Auto-SRL vs. Unregulated AI Tutor)	Cat D	Within-AI Closed Paper Exam (W-04a, W-04b)
Cosentino et al. (2025)	Ages 11–13, Norway (N = 30 analysed / 34 enrolled)	Lab-in-School RCT (GPT-4 vs. experienced math teacher on MOVES-NL)	Cat A	Active Human Benchmark (H-02)

Report	Sample & Setting	Design & Generative AI Modality	Cat.	Synthesis Role
Li et al. (2026)	Grades 6–8, USA (N = 386, 3-week 2×2×3 RCT)	Factorial Voice Design RCT inside ALTER-Math teachable agent	Cat B	Within-AI Modality Trial (ALTER-Math family)
J. Liu et al. (2025a)	Primary, China (n = 49 retained AI-arm pupils)	Quasi-Exp. report omitting control posttest (ChatGPT-MPS)	Cat A/B	Single-arm paired analysis (ChatGPT-MPS family)
J. Liu et al. (2025b)	Primary, China (n = 49 paired pupils)	GCCCE conference companion reporting identical n = 49 score block	Cat A/B	Linked companion of J. Liu et al. (2025a) (ChatGPT-MPS family)
J. Liu et al. (2025c)	Primary, China (pupils & 3 teachers)	Cross-sectional usability survey & teacher interviews (ChatGPT-MWPS)	Aux	Process & engagement synthesis
Z. Liu et al. (2025)	Grades 6–8, USA (N = 589 paired pupils)	Single-group pre–post + gain-selected strata log analysis (ALTER-Math)	Cat B	Process log synthesis (linked to Xing 2025)
Öner (2025)	Grade 6, Turkey (N = 13)	Single-group pre–post activity on arithmetic mean with ChatGPT	Cat B	Single-arm descriptive synthesis
Bulut & Borromeo Ferri (2026)	Secondary, Germany (dyadic pairs)	Qualitative observational study of mathematical modelling & verification	Cat A	Qualitative process synthesis
Rizos et al. (2024)	Grade 9 SEN, Greece (2 focal pupils in class)	Qualitative classroom case with ChatGPT-generated differentiated sheets	Cat C	Qualitative SEN synthesis (no-teacher-help tests)
Walkington et al. (2025)	Middle School, USA (N = 36 girls, residential camp)	Qualitative interaction analysis of algebra problem posing with AI	Aux	Qualitative problem-posing synthesis
Besas, Patac, & Patac (2026)	Senior High, Philippines (survey + focus group)	Cross-sectional self-directed ChatGPT use survey & school math grades	Obs	Observational & qualitative synthesis
Nguyen et al. (2026)	Primary School, Vietnam (2 intact classes)	AI-assisted cultural storytelling geometry media (statcheck discrepancy)	Aux/Flag	Static teacher-media boundary comparator
Trocado et al. (2026)	Grade 10, Portugal (N = 26 in pairs)	Qualitative exploration of quadratic functions supported by ChatGPT	Cat A	Qualitative process synthesis
Norberg et al. (2025)	Middle School, USA (N = 72,905 in MATHia)	Embedded RCTs of offline GPT-4 readability rewrites of word problems	Cat A/Aux	Static offline text-simplification boundary
Baez et al. (2025)	Secondary, USA/LatAm (12 group artefacts)	Educator preference ratings of AI-assisted polynomial art projects	Aux	Educator artefact rating synthesis
Kuzu (2026)	Primary, Germany (multilingual small groups)	Qualitative interpretative video analysis on translanguaging & math	Cat A	Qualitative multilingual synthesis
Karatayeva et al. (2026)	Grade 4, Kazakhstan (N = 263, 10 classes)	Matched Cluster Quasi-Exp. of static Gemini 2.5 Pro paper handouts	Cat B/Aux	Static paper-handout boundary comparator
Lintner et al. (2026)	Grade 7, Czechia (N = 2,307 nationwide)	Nationwide cross-sectional survey of GenAI use, critical thinking & math	Obs	Large-scale observational synthesis
X. Wang & Wei (2025)	Grade 6, China (N = 105)	Quasi-Exp. on mathematics anxiety (MARS scale; no attainment test)	Aux	Affective outcome synthesis
Ergene & Çaylan Ergene (2025)	High School, Turkey (160 pupils, 80 teachers)	Comparative rating of 5 GenAI solvers across 390 e-textbook tasks	Aux	Epistemic evaluation synthesis
Serin & Kepceoğlu (2026)	Grade 7, Turkey (N = 40, 2 intact classes)	Classroom exploratory comparison of GPT-5 Hint vs. Critique prompting	Cat A	Problem-posing uptake synthesis
Dello Iacono & Picariello (2026)	Grade 13, Italy (N = 12, two sessions)	Qualitative metacognitive coding of function derivative tasks with ChatGPT	Cat A	Qualitative calculus metacognition synthesis
Y. Kim, Lee, Shin, & Park (2025)	Primary School, South Korea (case study)	Commognitive discourse case study in AI levelling-off mean environment	Cat A	Qualitative commognitive synthesis
N. Kim & Chang (2026)	Grade 5, South Korea (N = 6)	Qualitative case study of real-world context integration in modelling	Cat A	Qualitative modelling synthesis
So, Jung, Seong, & Kim (2026)	Grade 10, South Korea (N = 3)	Commognitive analysis of roller-coaster curve modelling with ChatGPT	Cat A	Qualitative commognitive synthesis
Lee, Park, Shin, & Kwak (2025)	Grade 9, South Korea (N = 6)	Qualitative study of statistical refutation & counter-arguments on box plots	Cat A	Qualitative statistical reasoning synthesis

Report	Sample & Setting	Design & Generative AI Modality	Cat.	Synthesis Role
Yunianto, Galiç, & Lavicza (2024)	Lower Secondary, Indonesia (classroom EDR)	Educational design research integrating ChatGPT + GeoGebra for CT	Cat A	Qualitative geometry/CT synthesis
Contel & Cusi (2025)	Upper Secondary, Italy (classroom pilot)	Qualitative study of metacognitive utilisation schemes with GPT-4	Cat A	Qualitative metacognitive synthesis
Canonigo (2026)	Secondary, Philippines (interpretivist study)	Qualitative analysis of teacher critical mediation, agency & equity	Aux	Qualitative equity & agency synthesis
Oğuz Haçat & Kepceoğlu (2025)	Grade 7, Turkey (N = 4)	Exploratory case study of ChatGPT on geographic/demographic math skills	Cat A	Qualitative interdisciplinary math synthesis

Note. Cat. = Assessment-condition category (A = GenAI active during test; B = posttest with unresolved resource rules; C = partial human-help restriction; D = explicit closed-resource test; Obs = observational cross-sectional; Aux = affective, rating, or static/offline boundary outcome). Linked reports from the same cohort (J. Liu et al., 2025a, 2025b, 2025c; Xing et al., 2025; Li et al., 2026; Z. Liu et al., 2025) are explicitly cross-referenced and never double-counted.

3.2 Assessment-Stratified Meta-Analysis: Supported Performance Versus Post-Withdrawal Attainment (RQ1)

Table 2 and Figure 2 report the primary Tier-1 synthesis (Model 1, $k = 14$ contrasts across 8 multi-cluster or individually randomized studies) and the extended controlled synthesis (Model 2, $k = 28$ contrasts across all 17 GenAI-vs.-non-AI studies). Pooled effects fall step by step as test-time access to external help closes.

Assisted testing inflates apparent mastery. Live GenAI access during testing (Category A) produces large, statistically significant gains: g_PM = +0.615 (95% CI [+0.044, +1.185], $p = .035$, $I^2 = 96.6\%$) across primary trials (Bastani et al., 2025; Oreopoulos et al., 2026) and g_PM = +0.542 ([+0.118, +0.965], $p = .012$, $I^2 = 94.9\%$) when adding the 11-quarter ALEKS natural experiment of Rismanchian et al. (2026). Wide dispersion within Category A mirrors the scope of live assistance permitted by each interface. Step-by-step hints from GPT Tutor in Bastani et al. (2025) boosted practice accuracy by $g = +1.181$ (+127% over textbook practice) and unconstrained GPT Base raised scores by $g = +0.485$ (+48%), whereas post-error walkthroughs in Oreopoulos et al. (2026) lifted subsequent practice-set performance far more modestly ($g = +0.202$).

Unverified posttests shrink that advantage. Immediate posttests with undocumented resource rules (Category B) cut the primary pooled effect by over two-thirds to g_PM = +0.186 ([+0.009, +0.362], 95% HKSJ CI [−0.064, +0.435], $p = .039$, $I^2 = 0.0\%$, g_REML = +0.209) across elementary Socratic Pólya scaffolding (Kuo et al., 2025), middle-school teachable-agent work (Xing et al., 2025), context-personalized word problems (Tasdelen & Bodemer, 2025), and static LLM explanations (Lademann et al., 2025); adding Tier-2 intact-classroom studies across all 17 Category B contrasts yields g_PM = +0.326 ([+0.081, +0.571], $p = .009$, $I^2 = 51.2\%$). Partial human-help restriction (Category C) appears in primary models via the 900-tutor RCT of Tutor CoPilot (R. E. Wang et al., 2025), where tutors refrained from helping on session exit tickets and unconditional passing rose by 4.0 percentage points ($g = +0.095$ [+0.024, +0.167], $p = .009$).

Closed-resource exams tell a starkly different story. On explicit Category D immediate and delayed assessments barring GenAI and unauthorized aids, the pooled primary estimate ($k = 5$ contrasts across Bastani et al., 2025; Oreopoulos et al., 2026; and the 8-month invigilated paper endline of Henkel et al., 2024) sits right at zero: g_PM = −0.010 (95% CI [−0.137, +0.117], HKSJ [−0.190, +0.170], $p = .876$, $I^2 = 67.1\%$, g_REML = −0.011). Adding the proctored ALEKS retention checks of Rismanchian et al. (2026) (OR = 0.75, $g = −0.159$) in Model 2 ($k = 6$) gives an equally null g_PM = −0.035 ([−0.149, +0.078], HKSJ [−0.185, +0.114], $p = .543$, $I^2 = 68.0\%$). Omnibus moderation across categories is significant under common-τ^2 mixed-effects testing in primary trials (Q_M(3) = 9.17, $p = .027$; separate-τ^2 Q_between(3) = 6.78, $p = .078$) and strong across the extended corpus (Q_between(3) = 12.55, $p = .0059$). Our pre-specified focal contrast

between Supported Performance (Category A) and Closed-Resource Performance (Category D) confirms post-withdrawal attenuation of $\Delta g = +0.625$ ($z = 2.10$, $p = .036$) in primary trials and $\Delta g = +0.577$ ($z = 2.58$, $p = .010$) overall.

Table 2. *Assessment-Stratified Random-Effects Meta-Analysis of GenAI Versus Non-AI Mathematics Instruction and Sensitivity Analyses*

Model & Assessment Stratum	k	g_PM	95% Wald CI	95% HKSJ CI	I^2 (τ^2_PM)	p (g_REML)
Model 1: Primary Tier-1 Corpus	**14**	—	—	—	—	—
Category A: Supported performance (GenAI active)	3	+0.615	[+0.044, +1.185]	[−0.638, +1.867]	96.6% (0.243)	.035 (+0.615)
Category B: Posttest, resource rules unresolved	5	+0.186	[+0.009, +0.362]	[−0.064, +0.435]	0.0% (0.000)	.039 (+0.209)
Category C: Partial human-help restriction	1	+0.095	[+0.024, +0.167]	—	0.0% (—)	.009 (+0.095)
Category D: Explicit closed-resource exam / retention	5	−0.010	[−0.137, +0.117]	[−0.190, +0.170]	67.1% (0.015)	.876 (−0.011)
Model 1 Moderation ($Q_M(3) = 9.17$, $p = .027$; $Q_{between}(3) = 6.78$, $p = .078$) & Focal A vs. D (Δg)	—	+0.625	z = 2.10	—	—	.036
Model 2: Extended Tier-1 + Tier-2 Corpus	**28**	—	—	—	—	—
Category A: Supported performance (GenAI active)	4	+0.542	[+0.118, +0.965]	[−0.146, +1.229]	94.9% (0.177)	.012 (+0.542)
Category B: Posttest, resource rules unresolved	17	+0.326	[+0.081, +0.571]	[+0.061, +0.591]	51.2% (0.130)	.009 (+0.324)
Category C: Partial human-help restriction	1	+0.095	[+0.024, +0.167]	—	0.0% (—)	.009 (+0.095)
Category D: Explicit closed-resource exam / retention	6	−0.035	[−0.149, +0.078]	[−0.185, +0.114]	68.0% (0.014)	.543 (−0.034)
Model 2 Moderation ($Q_{between}(3) = 12.55$, $p = .0059$; $Q_M(3) = 7.36$, $p = .060$) & Focal A vs. D (Δg)	—	+0.577	z = 2.58	—	—	.010
Sensitivity Grids on Primary Model 1	—	—	—	—	—	—
Sens. A1: Oreopoulos 1-wk test in Cat C — Category C	3	+0.039	[−0.049, +0.128]	[−0.155, +0.233]	80.0% (0.005)	.387 (+0.039)
Sens. A2: Oreopoulos 1-wk test in Cat C — Category D	3	−0.009	[−0.292, +0.275]	[−0.630, +0.613]	65.0% (0.046)	.953 (−0.009)
Sens. B1: Lower classroom ICC (ρ = 0.050) — Cat D (Cat B = +0.222)	5	+0.010	[−0.145, +0.164]	[−0.205, +0.224]	$Q_M = 8.51$	.036
Sens. B2: Higher classroom ICC (ρ = 0.200) — Cat D (Cat B = +0.172)	5	−0.017	[−0.129, +0.096]	[−0.177, +0.144]	$Q_M = 9.62$	.022
Sens. C: Leave-One-Out range across Primary Category D	4	−0.036 to +0.019	All 95% CIs span 0	—	47.5%–75.3%	≥ .596

Note. g_PM = Paule–Mandel random-effects pooled Hedges' g adjusted for classroom/school clustering at empirical ICC ρ = 0.128 where primary studies reported unclustered statistics; g_REML = restricted maximum likelihood point estimate. HKSJ = Hartung–Knapp–Sidik–Jonkman small-sample t-interval.

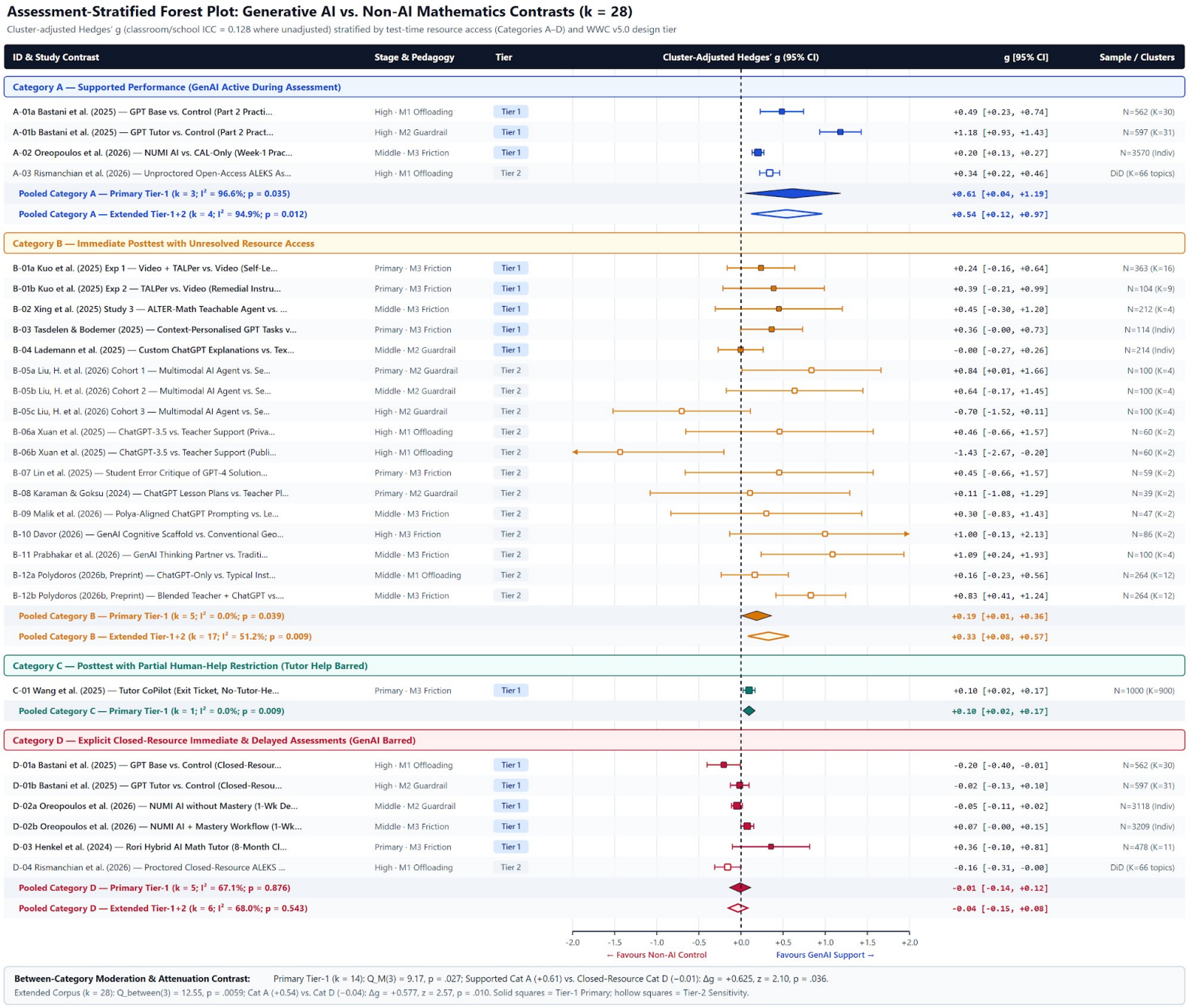


Figure 2. Assessment-Stratified Forest Plot Across All 28 GenAI-Versus-Non-AI School-Mathematics Contrasts (Categories A–D) Note. Solid squares and filled diamonds denote Tier-1 primary multi-cluster or individually randomised contrasts (k = 14); open squares and hollow-fill diamonds denote Tier-2 sensitivity contrasts (k = 14) and extended subgroup totals (k = 28).

Three sensitivity checks test this closed-resource null (Table 2, bottom panel). Moving the Week-2 teacher-invigilated test of Oreopoulos et al. (2026) from Category D to Category C leaves both strata near zero (Category C: g = +0.039; Category D: g = −0.009). Shifting imputed classroom clustering between ρ = 0.05 and ρ = 0.20 keeps Category D within +0.010 to −0.017 and preserves omnibus moderation (p = .036 to .022). Leave-One-Out refitting across Category D shows that dropping either the negative GPT Base exam (D-01a; remaining g = +0.019, p = .711) or the positive 8-month Rori paper endline (D-03; remaining g = −0.027, p = .596) leaves the null intact.

3.3 Pedagogical Guardrails, Enforced Productive Friction, and Within-AI Trials (RQ2)

Why do post-withdrawal mathematics estimates span severe learning loss (g = −0.205 in Bastani GPT Base; g = −1.435 in Xuan public high school) alongside steady positive gains (g = +0.095 to +0.453 in structured tutoring workflows)? Table 3 and Figure 3A evaluate the two-threshold pedagogical model across all post-withdrawal GenAI-vs.-non-AI contrasts (Categories B, C, and D).

Pedagogical architecture explains 100% of between-contrast variance in primary Tier-1 trials (Model 3a, k = 11; I^2 = 0.0% within every stratum; Q_between(2) = 18.69, p = .0001):

- *M1: Direct-Answer Offloading (*k *= 1 primary;* k *= 5 extended). Direct answers hurt retention. Unconstrained GPT-4 Base access during high-school practice depressed subsequent closed-exam scores (*g *= −0.205 [−0.403, −0.007],* p *= .043; Bastani et al., 2025). Pooling the extended M1 contrasts, which add unproctored-to-proctored offloading in ALEKS (*g *= −0.159; Rismanchian et al., 2026), unguided ChatGPT-3.5 in Vietnamese high schools (Xuan et al., 2025), and unmediated ChatGPT-only algebra in Greece (*g *= +0.164, n.s.; Polydoros, 2026), yields a negative, non-significant average (*g *= −0.148 [−0.607, +0.311],* p* = .527).*
- *M2: Passive Hint Guardrails and Static Explanations (*k *= 3 primary;* k *= 7 extended). Passive guardrails halt the bleeding. System prompts that withhold final answers (Bastani et al., 2025, GPT Tutor:* g *= −0.015), conversational post-error walkthroughs without required mastery streaks (Oreopoulos et al., 2026, Non-Mastery:* g *= −0.048), and custom ChatGPT proportional-reasoning explanations (Lademann et al., 2025:* g *= −0.004) eliminated M1 harm (Threshold 1) but delivered zero net gain over standard non-AI instruction (Primary* g *= −0.038 [−0.093, +0.017],* p *= .173,* I^2 *= 0.0%; Extended* g *= +0.054 [−0.239, +0.346],* p *= .720). Lademann et al. (2025) captured this fluency illusion cleanly: sixth-graders reading ChatGPT explanations reported lower cognitive load (*d *= −0.59), higher positive emotion (*d *= +0.48), and higher self-efficacy (*d *= +0.32) than textbook peers while scoring identically on the posttest (50.29% vs. 50.38%,* p* = .75).*
- *M3: Enforced Productive Friction (*k *= 7 primary;* k *= 12 extended). Durable independent transfer emerged only when instructional designs demanded active learner generation (Threshold 2): Primary* g *= +0.100 [+0.050, +0.151],* p *= .0001,* I^2 *= 0.0%; Extended* g *= +0.315 [+0.150, +0.481],* p *= .0002 (Model 3b* Q_*M(2) = 8.44,* p *= .015). Productive friction took five concrete forms: 3-in-a-row unaided mastery gating paired with post-error AI walkthroughs (Oreopoulos Mastery: +3.2 pp overall and +5.9 pp,* p *= .017 on hard items one week later; Oreopoulos et al., 2026), Socratic questioning tied to Pólya four-step planning in TALPer (+6.41 points for low-achieving fifth-graders in self-study and +15.06 points,* p *= .017, in remediation; Kuo et al., 2025), teaching a fallible AI peer in ALTER-Math (Xing et al., 2025) or debugging planted GPT-4 errors (Lin et al., 2025), structured micro-lesson scaffolding in Rori (Henkel et al., 2024), and live pedagogical coaching of human tutors in Tutor CoPilot (R. E. Wang et al., 2025). Cross-study subgroups can still differ in student intake or curriculum. Model 3c therefore pools the six head-to-head within-AI experiments (W-01 to W-05) in which both randomized arms used GenAI under contrasting pedagogical rules (Table 3, lower panel): across the five peer-reviewed and working-paper primary contrasts, guardrails and metacognitive or teacher regulation beat unguarded AI by* g_*PM = +0.658 ([+0.123,*

+1.192], p = *.016), and across all six contrasts (*k = *6) the pooled gain is* g_*PM = +0.645 ([+0.215, +1.075], HKSJ [+0.081, +1.209],* p* = .0033).

Week-7 closed-paper testing in Luo et al. (2026) shows this mechanism in sharp relief. Fifth-graders solved five Pólya-scored equation word problems without AI. Unregulated AI MathTutor access during training invited rampant answer copying and left median paper performance at just 4.0 / 40, whereas pairing that same tutor with pre/post-task self-regulated learning (SRL) chatbots doubled unaided paper scores (Median = 8.0, g = +0.846, p = .002) and adding human-in-the-loop (HITL) teacher review of delegation logs curbed direct solution requests (r = 0.59, p = .008) and tripled paper scores (Median = 13.0, g = +1.676, p < .001). Worden et al. (2025) observed a parallel split in middle school: replacing bare answers with self-verified GPT-4 step-by-step explanations quadrupled the odds of solving the next item unassisted (Exp 1: OR = 4.18, g = +0.789, p < .001) and matched expert teacher explanations to within four-hundredths of a standard deviation (Exp 2, N = 6,371: g = +0.006 [−0.033, +0.044], p = .740; pooled Active Human Benchmarks Model 3d with Cosentino et al., 2025: g = +0.005 [−0.034, +0.043], p = .811, I^2 = 0.0%).

Table 3. *Pedagogical Architecture Moderation of Post-Withdrawal Mathematics Outcomes, Head-to-Head Within-AI Trials, and Human Benchmarks*

Model, Pedagogical Tier, or Within-AI Experimental Contrast	k	Hedges' g	95% CI	I²	p
Model 3a: Primary Post-Withdrawal Pedagogical Moderator (Cat B+C+D)	**11**	—	—	—	—
M1: Direct-Answer Offloading (unconstrained LLM)	1	−0.205	[−0.403, −0.007]	0.0%	.043
M2: Passive Hint Guardrails / Static AI Explanations	3	−0.038	[−0.093, +0.017]	0.0%	.173
M3: Enforced Productive Friction (Mastery / Teachable / Pólya / CoPilot)	7	+0.100	[+0.050, +0.151]	0.0%	.0001
Model 3a Omnibus Moderation: Separate-τ² Q_between(2) = 18.69; Common-τ² Q_M(2) = 18.69	—	—	—	—	.0001
Model 3b: Extended Post-Withdrawal Pedagogical Moderator (Cat B+C+D)	**24**	—	—	—	—
M1: Direct-Answer Offloading	5	−0.148	[−0.607, +0.311]	50.1%	.527
M2: Passive Hint Guardrails / Static AI Explanations	7	+0.054	[−0.239, +0.346]	39.3%	.720
M3: Enforced Productive Friction	12	+0.315	[+0.150, +0.481]	56.5%	.0002
Model 3b Omnibus Moderation: Common-τ² Q_M(2) = 8.44 (p = .015); Separate-τ² Q_between(2) = 4.98	—	—	—	—	.081
Model 3c: Direct Head-to-Head Within-AI Guardrail Contrasts (Both Arms Used GenAI)	—	—	—	—	—
W-01 Bastani et al. (2025): GPT Tutor (Hint) vs. GPT Base (Closed Exam)	1	+0.193	[+0.017, +0.370]	—	.038
W-02 Oreopoulos et al. (2026): AI + Mastery vs. AI without Mastery (1-Wk)	1	+0.123	[+0.023, +0.223]	—	.016
W-03 Worden et al. (2025) Exp 1: GPT-4 CoT Explanation vs. Direct Answer	1	+0.789	[+0.373, +1.204]	—	<.001
W-04a Luo et al. (2026): Auto-SRL Bots + AI vs. Unregulated AI (Closed Paper)	1	+0.846	[+0.228, +1.464]	—	.002
W-04b Luo et al. (2026): Teacher HITL + SRL + AI vs. Unregulated AI (Paper)	1	+1.676	[+0.980, +2.371]	—	<.001
W-05 Polydoros (2026, Preprint): Blended Teacher + AI vs. ChatGPT-Only	1	+0.663	[+0.329, +0.997]	—	<.001
Pooled Within-AI Guardrail Advantage — Primary Trials (W-01 to W-04b)	5	+0.658	[+0.123, +1.192]	87.3%	.016

Model, Pedagogical Tier, or Within-AI Experimental Contrast	k	Hedges' g	95% CI	I^2	p
Pooled Within-AI Guardrail Advantage — All Trials (W-01 to W-05)	6	+0.645	[+0.215, +1.075]	87.0%	.0033
Model 3d: Active Expert Human Benchmarks (GPT-4 vs. Experienced Math Teacher)	—	—	—	—	—
Pooled Human-Equivalence Benchmark (Worden et al., 2025, Exp 2; Cosentino et al., 2025)	2	+0.005	[−0.034, +0.043]	0.0%	.811

Note. In Luo et al. (2026), effective control sample size (CG n = 32 → n_eff = 16 per contrast) was split across W-04a and W-04b in pooled Model 3c to prevent double-counting. HKSJ 95% CI for Model 3c (k = 5) is [−0.099, +1.414] and for Model 3c (k = 6) is [+0.081, +1.209].

3.4 Developmental Stage Moderation and Senior-Secondary Vulnerability (RQ3a)

School level moderated post-withdrawal outcomes in both primary trials (Model 4a: Q_between(2) = 6.35, p = .041; Q_M(2) = 7.29, p = .026) and the extended pool (Model 4b: Q_M(2) = 6.76, p = .033; Figure 3B). Elementary cohorts (Grades 3–5) posted steady positive transfer (Primary Tier-1 k = 5: g = +0.128 [+0.046, +0.210], p = .002; Extended k = 8: g = +0.140 [+0.055, +0.225], p = .0012, I^2 = 6.2%), while middle-school cohorts (Grades 6–8) ranged from null to moderately positive depending on mastery or teachable-agent rules (Primary k = 4: g = +0.015 [−0.072, +0.101]; Extended k = 9: g = +0.272 [+0.025, +0.518], p = .031). Upper-secondary cohorts reversed sign. High-school interventions (Grades 9–12) trended negative in both primary (k = 2: g = −0.091 [−0.273, +0.091]) and extended syntheses (k = 7: g = −0.149 [−0.616, +0.317], p = .530, I^2 = 60.1%). Two distinct pressures converge in upper-secondary classrooms. Older students enjoy broad device autonomy that invites rapid symbolic offloading: in Rismanchian et al. (2026), high-schoolers cut ALEKS study time by 31.3% after ChatGPT launched, pushed unproctored accuracy odds up by 85%, and lost 25% in proctored retention odds. Multi-step calculus and limit derivations also tax working memory heavily when delivered as dense chatbot text. H. Liu et al. (2026) tested an identical multimodal LLAMA-13B tutor across three school stages and found large gains in Grade 5 fractions (g_adj = +0.836) and Grade 8 Pythagorean geometry (g_adj = +0.637), yet Grade 11 derivative comprehension fell sharply (−9.43 points, g_adj = −0.704) as slide dwell time dropped by half (8.6 s vs. 17.2 s) under opaque symbolic output. Xuan et al. (2025) documented a similar divergence on Grade 11 limits: unguided ChatGPT-3.5 modestly helped ICT-fluent private-school pupils (g_adj = +0.457) but triggered severe posttest collapse (g_adj = −1.435) in an under-resourced public school despite higher baseline IQ in the AI classroom (98.3 vs. 91.4, p < .01).

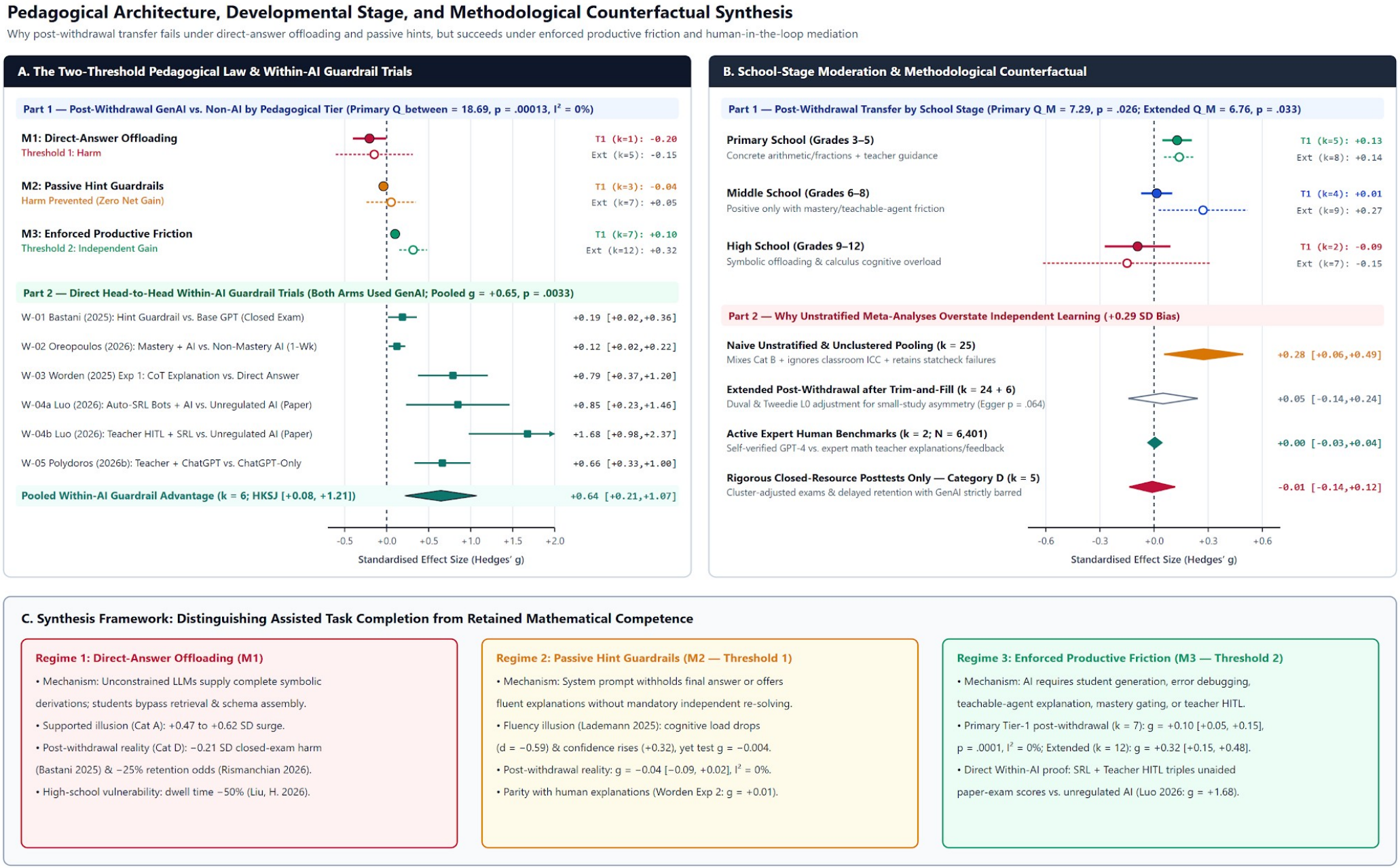


Figure 3. Two-Threshold Pedagogical Moderation, Head-to-Head Within-AI Trials, Developmental Stage Effects, and Methodological Counterfactual Note. Panel A contrasts between-group pedagogical tiers (M1–M3) with direct head-to-head within-AI guardrail trials (W-01 to W-05). Panel B displays school-stage moderation alongside the +0.29 SD methodological bias of naive unstratified synthesis. Panel C summarises the three operational regimes.

3.5 Qualitative Process Synthesis, Observational Evidence, and Boundary Comparators (RQ3b)

Task framing decides how GenAI behaves in classrooms. Across the 28 non-pooled studies and embedded process logs, unmediated student prompting routinely slid into procedural shortcutting, uncritical copying of hallucinated algebraic or geometric steps, and shallow context borrowing (Bulut & Borromeo Ferri, 2026; Dello Iacono & Picariello, 2026; N. Kim & Chang, 2026; Trocado et al., 2026; Walkington et al., 2025; Yunianto et al., 2024). Mathematical agency returned when teachers cast the model as a fallible partner whose claims demanded proof. Tenth-graders in Contel and Cusi (2025) sharpened metacognitive monitoring by auditing flawed ChatGPT algebraic proofs, ninth-graders in Lee et al. (2025) defended box-plot interpretations against AI counterarguments, and seventh-graders given GPT-5 critique prompts revised problem-posing tasks more deeply than peers given direct hints (Serin & Kepceoğlu, 2026). Commognitive analyses traced object-level narrative growth as pupils negotiated visual leveling-off models of the arithmetic mean (Y. Kim et al., 2025), polynomial curve parameters (So et al., 2026), multilingual Fermi estimates (Kuzu, 2026), map-scale ratios (Oğuz Haçat & Kepceoğlu, 2025), and adapted SEN worksheets (Rizos et al., 2024). Turkish high-schoolers and teachers rating five AI solvers in Ergene and Çaylan Ergene (2025) and educator judges in Baez et al. (2025) likewise tied instructional worth to representational clarity rather than raw answer speed, while Canonigo (2026) documented socioeconomic gaps in home AI reliance.

Uncontrolled pre–post gains cannot stand in for Category D trials. Five single-group or gain-stratified studies illustrate that pitfall clearly. Linked ChatGPT-MPS papers reported paired gains among 49 elementary pupils while omitting posttest data from the comparison class (J. Liu et al., 2025a, 2025b; usability companion J. Liu et al., 2025c). Log analyses of 589 middle-schoolers in ALTER-Math (Z. Liu et al., 2025) showed 55.5% improving and 35.0% declining on identical 15-item pre/post tests; contrasting engagement logs across post-hoc gain groups isolates behavioral correlates rather than treatment effects, while a companion factorial RCT in ALTER-Math (N = 386; Li et al., 2026) showed voice personalization lifting social presence. Small single-arm gains on sixth-grade mean items (N = 13; Öner, 2025) and math-anxiety drops without achievement testing (N = 105; X. Wang & Wei, 2025) likewise leave independent retention untested. Two naturalistic homework surveys point the same way: senior-high ChatGPT frequency in Besas et al. (2026) and nationwide Czech seventh-grade GenAI uptake in Lintner et al. (2026; N = 2,307) showed neutral-to-negative links with school math grades unless paired with habit-level verification.

Three static/offline boundary studies confirm that using an LLM behind the scenes to prepare fixed materials differs fundamentally from live student–AI chat. Offline GPT-4 readability rewrites of MATHia word problems across 72,905 pupils left workspace promotion unchanged overall while trimming hint requests slightly (Norberg et al., 2025). Teacher-vetted Gemini 2.5 Pro paper worksheets raised Grade 4 procedural scores because pupils had zero live offloading channel (Karatayeva et al., 2026), whereas AI-illustrated geometry storytelling in Vietnamese primary classes (Nguyen et al., 2026) mixed teacher delivery with internal t/SD discrepancies catalogued in Supplement S1 (§S1.3).

3.6 Risk of Bias, Small-Study Asymmetry, and the Methodological Counterfactual (RQ4)

Trial quality split cleanly along design lines. Outcome-level quality appraisal (Supplement S1, §S1.5) assigned low-to-moderate risk of bias only to multi-cluster or individually randomized trials with verified closed-resource or platform-blinded endpoints (Bastani et al., 2025; Lademann et al., 2025; Oreopoulos et al., 2026; Tasdelen & Bodemer, 2025; Worden et al., 2025; rated Some concerns where open classroom delivery precluded student blinding). Nearly every Category B intact-classroom quasi-experiment carried serious risk of bias from single-cluster-per-arm teacher confounding (Davor, 2026; Karaman & Göksu, 2024; Lin et al., 2025; Malik et al., 2026; Xuan et al., 2025), teacher self-selection into treatment arms (Kuo et al., 2025), or unrecorded exam proctoring.

Funnel diagnostics show how unverified Category B trials pull naive averages upward. Across the extended post-withdrawal set (k = 24), Egger's regression signaled borderline small-study asymmetry (b_0 = +0.834, SE = 0.450, z = 1.85, p = .064), and Duval and Tweedie's L0 Trim-and-Fill imputed k_0 = 6 left-tail contrasts and reduced the pooled estimate from g = +0.174 [+0.008, +0.340] to a non-significant g_TF = +0.050 [−0.142, +0.242]. Primary Tier-1 post-withdrawal trials (k = 11) showed no funnel asymmetry (Egger p = .236; g = +0.054 [−0.044, +0.151]; g_TF = +0.025).

Model 5 prices out the exact methodological counterfactual (RQ4, Figure 3B). Pooling post-intervention mathematics outcomes without assessment-condition stratificationwhile leaving classroom ICC clustering unadjusted and retaining an unscreened statcheck-flagged posttest alongside the 24 post-withdrawal contrasts (k = 25) produces an apparent headline gain of g = +0.275 (95% CI [+0.056, +0.494], p = .014, I^2 = 91.6%). Set against verified restricted posttests (Categories C + D: g = +0.012 [−0.093, +0.117], p = .826) or strict closed-resource exams (Category D: g = −0.010 [−0.137, +0.117], p = .876), naive synthesis inflates independent mathematics learning by +0.26 to +0.29 SD.

4. Discussion

4.1 The Two-Threshold Framework of Independent Mathematical Transfer

Synthesizing all 49 reports resolves the apparent clash between large assisted practice gains and null or negative exam transfer. Two principles govern the evidence. Assessment rules define what a score measures: live GenAI access (Category A, g = +0.54 to +0.62) captures human–AI co-performance, whereas closed examinations (Category D) strip away the tool and reveal a zero average residue (g = −0.01 to −0.04). Post-withdrawal learning then follows a Two-Threshold Pedagogical Law (Figure 3C):

- **Threshold 1 (Blocking Direct Offloading Stops Harm, Not Stagnation):** Moving from unconstrained answer generation (M1, g = −0.21) to prompt-level hint guardrails or static LLM explanations (M2, g = −0.04) eliminates copy-and-paste skill erosion and matches expert human hints (g = +0.005). Because fluent walkthroughs lower perceived effort without forcing independent retrieval, passive guardrails still leave net transfer at zero.
- **Threshold 2 (Enforcing Productive Cognitive Friction Builds Competence):** Positive post-withdrawal transfer (M3, g = +0.10 in large primary trials; g = +0.32 across extended classrooms; g = +0.65 in head-to-head within-AI trials) requires workflows that keep the core mathematical work on the student through consecutive unaided mastery streaks, Socratic Pólya planning, fallible-agent teaching or error debugging, real-time coaching of human tutors, or teacher review of AI delegation logs.

4.2 Design and Reporting Standards for School-Mathematics AI Research

Seventeen of the 24 post-withdrawal contrasts omitted basic test-security details (Category B). Cumulative progress in K–12 AI research requires stricter trial reporting, so Table 4 sets out eight minimum design and disclosure standards for future evaluations.

Table 4. *Minimum Design and Reporting Standards for Evaluating Independent Mathematics Learning After Generative AI Support*

Standard	Operational Requirement for Primary Trials and Peer Review
1. Explicit Test-Time Resource Audit	State verbatim whether GenAI, calculators, web search, notes, peer talk, and teacher hints were barred, monitored, or permitted during every posttest.
2. Separation of Supported vs.	Report performance *during* AI-supported practice (Category A) separately from

Standard	Operational Requirement for Primary Trials and Peer Review
Unaided Estimands	invigilated *closed-resource* posttests (Category D); never average them.
3. Delayed Retention Beyond 24 Hours	Include at least one closed-resource assessment ≥ 1–4 weeks after AI withdrawal on non-identical transfer items (not exact practice repeats).
4. Pedagogical Guardrail & Friction Specification	Publish exact system prompts, answer-withholding rules, mastery-gating criteria, and logs quantifying direct answer requests vs. generative prompts.
5. Multi-Cluster Power & Variance Reporting	Avoid 1-class-per-arm (K = 2) designs; randomise ≥ 10–20 classrooms/schools or individuals and report empirical outcome ICCs and cluster-robust SEs.
6. Active Instructional Comparator Parity	Match total mathematics instructional time, practice item exposure, and teacher attention across GenAI and non-AI arms.
7. Pre-Specified Achievement & Stage Subgroups	Test baseline-achievement and grade-band interactions (e.g., Johnson–Neyman regions) given divergent effects between low-achieving primary and high-school calculus learners.
8. Open Microdata & Statcheck Verification	Deposit anonymised item/pupil-level data and verify internal arithmetic consistency across raw M/SD tables, ANCOVA estimates, and degrees of freedom.

Note. Adopted to eliminate Category B ambiguity, single-cluster confounding, and supported-versus-independent estimand conflation.

4.3 Limitations of the Evidence Base and of This Review

Three boundaries constrain interpretation. Although full-text retrieval reached 100% (0 unretrieved), verified Category D trials remain concentrated in five primary contrasts and one large natural experiment, and only two studies measured multi-day or multi-month closed retention: Henkel et al. (2024) over 8 months (where Rori tutoring accompanied extra study-hall time) and Oreopoulos et al. (2026) at 1 week. Null short-term effects under passive guardrails do not rule out cumulative gains over full-year curricula paired with mastery gating. Classroom ICCs also required empirical imputation (ρ = 0.128, checked across 0.05–0.20) when primary reports omitted variance components, and binary or rank conversions (Cox log-odds, Ruscio AUC) carry standard scale approximations. Single-investigator screening and coding likewise warrant independent replication despite scripted numerical verification.

5. Conclusion

Generative AI raises school-mathematics performance substantially while active (g = +0.54 to +0.62) yet leaves verified closed-resource competence unchanged on average once withdrawn (g = −0.01 to −0.04). Unconstrained answer access erodes independent high-school exam skill. Passive hint prompts prevent that erosion without breaking the fluency illusion. Durable mathematical competence after AI withdrawal demands architectures that enforce productive cognitive friction: students must generate steps, debug errors, teach concepts, and pass unaided mastery checks under curricular and teacher oversight.

Declarations

Availability of Data, Code, and Materials: Online Supplement S1 archives all database and API search strings, item-by-item dispositions for the 30 Stream-1 candidates and 14 Stream-2 subscription-access articles, the forensic numerical audit ledger, complete effect-size derivations across all 36 contrasts, sensitivity grids, and domain-level RoB 2 and JBI matrices. Standalone replication scripts (meta_stats_engine.py, run_final_meta_analysis.py, reproduce_meta_analysis.R) and machine-readable tables (All_36_Meta_Analytic_Contrasts.csv) accompany the submission package.

Funding: External funding was not received for this study from any agency in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interests: No competing financial or non-financial interests exist.

Ethics Approval and Consent to Participate: Not applicable; this study is a PRISMA 2020 systematic review and meta-analysis of published aggregate study data and involved no direct recruitment of human participants.

References

Baez, R., Sanchez, H., Sandeep, N., & Pllana, D. (2025). Creativity and critical thinking in the AI age: A high school perspective. *International Journal of Research in Education and Science*, *11*(2), 448–477. https://doi.org/10.46328/ijres.1290

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), Article e2422633122. https://doi.org/10.1073/pnas.2422633122 (Correction: *PNAS*, 2025, *122*, e2518204122; open data: https://github.com/obastani/GenAICanHarmLearning)

Besas, J., Patac, L., & Patac, A., Jr. (2026). AI-facilitated self-directed learning and mathematics performance: A mixed-methods study on ChatGPT use among Generation Z students. *International Journal of Contemporary Educational Research*, *13*(1), 15–30. https://doi.org/10.52380/ijcer.2026.13.1.822

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Boolzen, C., Kuhn, J., Flegr, S., Rott, E. M., Stausberg, N., & Küchemann, S. (2026). Evidence of impact and interpretational limits of generative AI in STEM education: A systematic review and meta-analysis on cognitive learning outcomes. *Artificial Intelligence Review*, *59*, Article 223. https://doi.org/10.1007/s10462-026-11665-9

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). John Wiley & Sons.

Bulut, N., & Borromeo Ferri, R. (2026). AI-powered learning: Exploring secondary school students' use of ChatGPT for solving mathematical modelling problems. *ZDM – Mathematics Education*, *58*, 993–1009. https://doi.org/10.1007/s11858-026-01803-6

Canonigo, A. M. (2024). Leveraging AI-powered completion-based feedback to enhance students' conceptual understanding of quadratic functions. *Journal of Computer Assisted Learning*, *40*(6), 2890–2904. https://doi.org/10.1111/jcal.13065

Canonigo, A. M. (2026). Teacher mediation and the contingent promise of generative AI in mathematics education. *Journal of Computer Assisted Learning*, *42*, Article e70307. https://doi.org/10.1002/jcal.70307

Chen, S. (2025). Application and effectiveness analysis of AI-assisted mathematical modeling in STEAM education. In *Proceedings of the 2025 2nd International Conference on Artificial Intelligence and Future Education (AIFE 2025)* (pp. 1–7). ACM. https://doi.org/10.1145/3785987.3786040

Contel, F., & Cusi, A. (2025). Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education*, *11*, 167–191. https://doi.org/10.1007/s40751-024-00164-7

Cosentino, G., Anton, J., Sharma, K., Gelsomini, M., Giannakos, M., & Abrahamson, D. (2025). Generative AI and multimodal data for educational feedback: Insights from embodied math learning. *British Journal of Educational Technology*, *56*, 1686–1709. https://doi.org/10.1111/bjet.13587

Cox, D. R. (1970). *The analysis of binary data*. Methuen.

Davor, I. (2026). Generative AI-supported instruction as a cognitive scaffold: Effects on senior high school students' geometric reasoning and proof construction. *Frontiers in Education*, *11*, Article 1869961. https://doi.org/10.3389/feduc.2026.1869961

Dello Iacono, U., & Picariello, M. (2026). Investigating the effect of high school students' interactions with ChatGPT on metacognitive activities in mathematical problem solving. *International Journal for Technology in Mathematics Education*, *33*(2), 75–82. https://doi.org/10.1564/tme_v33.2.05

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Eldridge, S., Campbell, M. K., Campbell, M. J., Drahota, A. K., Giraudeau, B., Reeves, B. C., Siegfried, N., & Higgins, J. P. T. (2021, March 18). *Revised Cochrane risk-of-bias tool for cluster-randomized trials: Additional considerations for cluster-randomized trials*. Cochrane Methods. https://www.riskofbias.info/welcome/rob-2-0-tool/rob-2-for-cluster-randomized-trials

Ergene, Ö., & Çaylan Ergene, B. (2025). AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers. *Education and Information Technologies*, *30*, 509–545. https://doi.org/10.1007/s10639-024-13121-z

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, *20*(12), 1771–1782. https://doi.org/10.1002/sim.791

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). *Effective and scalable math support: Experimental evidence on the impact of an AI-math tutor in Ghana* (arXiv:2402.09809). arXiv. https://doi.org/10.48550/arXiv.2402.09809

Higgins, J. P. T., Savović, J., Page, M. J., & Sterne, J. A. C. (Eds.). (2019, August 22). *Revised Cochrane risk-of-bias tool for randomized trials (RoB 2): Guidance for individually randomized parallel-group trials*. Cochrane Methods. https://www.riskofbias.info/welcome/rob-2-0-tool/current-version-of-rob-2

Joanna Briggs Institute. (2024a, March 26). Appendix 4.3: Critical appraisal checklist for quasi-experimental studies. In *JBI manual for evidence synthesis*. https://jbi-global.atlassian.net/wiki/spaces/MANUAL/pages/355829117/

Joanna Briggs Institute. (2024b, March 26). Appendix 7.5: Critical appraisal checklist for analytical cross-sectional studies. In *JBI manual for evidence synthesis*. https://jbi-global.atlassian.net/wiki/spaces/MANUAL/pages/355599424/

Joanna Briggs Institute. (2025, March 2). 3.3.11 Assessing methodological quality of qualitative studies. In *JBI manual for evidence synthesis*. https://jbi-global.atlassian.net/wiki/spaces/MANUAL/pages/863928570

Kamaghe, J. (2026). Generative AI as a partner for teachers in building personalised learning paths for students with ease in Tanzania. *Research in Learning Technology*, *34*. https://doi.org/10.25304/rlt.v34.3594

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

Karaman, M. R., & Göksu, İ. (2024). Are lesson plans created by ChatGPT more effective? An experimental study. *International Journal of Technology in Education*, *7*(1), 107–127. https://doi.org/10.46328/ijte.607

Karatayeva, T., Ussenova, A., Khan, N. N., & Daumova, B. (2026). Mathematics outcomes and creative metacognition in primary students: A quasi-experimental study using AI-generated explanations. *International Journal of Learning, Teaching and Educational Research*, *25*(10), Article 23. https://doi.org/10.26803/ijlter.25.10.23

Kim, N., & Chang, H. (2026). Changes in the types of dealing with the context of task in elementary school students' mathematical modeling through interaction with generative AI. *Journal of Educational Research in Mathematics*, *36*(3), 461–491. https://doi.org/10.29275/jerm.2026.36.3.461

Kim, Y., Lee, K. H., Shin, D., & Park, J. (2025). Design and implementation of an AI-integrated mathematics teaching–learning environment for fostering primary school students' reasoning about the arithmetic mean. *School Mathematics*, *27*(4), 921–946. https://doi.org/10.57090/sm.2025.12.27.4.921

Kuo, B. C., Bai, Z. E., & Lin, C. H. (2025). Developing an AI learning companion for mathematics problem solving in elementary schools. *Computers & Education*, *240*, Article 105463. https://doi.org/10.1016/j.compedu.2025.105463

Kuzu, T. E. (2026). AI-supported translanguaging processes in primary school: Empirical insights into ChatGPT's role in multilingual interactions. *Technology, Knowledge and Learning*. Advance online publication. https://doi.org/10.1007/s10758-026-09974-7

Lademann, J., Henze, J., & Becker-Genschow, S. (2025). Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables. *Physical Review Physics Education Research*, *21*, Article 010147. https://doi.org/10.1103/PhysRevPhysEducRes.21.010147

Lee, K. H., Park, J., Shin, D., & Kwak, J. (2025). An analysis of reasoning in AI-chatbot-supported box-plot problem solving: Middle school students' reasoning in the result-interpretation stage. *Journal of the Korean School Mathematics Society*, *28*(4), 505–526. https://doi.org/10.30807/ksms.2025.28.4.007

Li, H., Xing, W., Li, C., & Lyu, B. (2026). I hear you: Enhancing middle school mathematics engagement and learning outcomes through teachable agent voice design. *Computers & Education*, *254*, Article 105725. https://doi.org/10.1016/j.compedu.2026.105725

Lin, Y. F., Yang, E. F. Y., Wu, J. S., & Yeh, C. Y. C. (2025). Enhancing students' authentic mathematical problem-solving skills and confidence through error analysis of GPT-4 solutions. *Research and Practice in Technology Enhanced Learning*, *20*, Article 034. https://doi.org/10.58459/rptel.2025.20034

Lintner, T., Juhanák, L., Žáková, M., Plch, L., & Ropovik, I. (2026). Generative AI use and academic achievement: A cross-sectional study of Czech seventh-grade students. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-026-14002-3

Liu, H., Li, C., Liu, Z., Yang, S., Wang, H., & Yan, P. (2026). Designing multimodal human–computer interaction for K–12 learning: Research design, system architecture, and field evaluation of an AI teaching agent. *International Journal of Human–Computer Interaction*, *42*(18), 14976–15001. https://doi.org/10.1080/10447318.2026.2616662

Liu, J., Keane, T., Sun, D., Yang, Y., & Yang, Y. (2025c). Design and evaluation of ChatGPT-MWPS: An AI-enhanced learning system for improving primary students' mathematical word problem solving. *Smart Learning Environments*, *12*, Article 68. https://doi.org/10.1186/s40561-025-00419-9

Liu, J., Sun, D., Sun, J., Wang, J., & Yu, P. L. H. (2025a). Designing a generative AI enabled learning environment for mathematics word problem solving in primary schools: Learning performance, attitudes and interaction. *Computers and Education: Artificial Intelligence*, *9*, Article 100438. https://doi.org/10.1016/j.caeai.2025.100438

Liu, J., Sun, D., Zheng, Z., & Xu, Z. (2025b). 基于生成式人工智能的小学数学应用题学习环境设计及其研究 (Design and research of a generative AI-based learning environment for primary school mathematics word problems). In *GCCCE 2025 Chinese Conference Proceedings* (pp. 901–904). Global Chinese Society for Computers in Education. https://nlt.gcsce.net/ojs/index.php/GCCCE/article/view/739

Liu, Z., Xing, W., Ngo, B., Jiao, X., Jiang, S., & Li, C. (2025). Engagement patterns of middle school students with AI teachable agents in mathematics learning. *Scientific Reports*, *15*, Article 40971. https://doi.org/10.1038/s41598-025-24841-8

López-Pernas, S., Misiejuk, K., & Saqr, M. (2026). How AI-generated feedback hinders or helps learning: A heterogeneous TNA study of learning dynamics. *Journal of Computer Assisted Learning*, *42*(4), Article e70285. https://doi.org/10.1002/jcal.70285

Luo, X., Wang, S., & Hew, K. F. (2026). Equipping elementary school students with self-regulated learning through human–AI collaboration in online learning. *Computers & Education*, *253*, Article 105697. https://doi.org/10.1016/j.compedu.2026.105697

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901–918. https://doi.org/10.1037/a0037123

Malik, M. A., Lince, R., & Husnaeni, H. (2026). A Pólya-aligned prompting protocol for ChatGPT scaffolding: Evidence from eighth-grade systems-of-linear-equations problem solving. *International Journal of Environment, Engineering and Education*, *8*(1), Article 331. https://doi.org/10.55151/ijeedu.v8i1.331

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Nguyen, H. H., Pham, S. N., Trinh, C. S., Dao, C. L. A., Nguyen, T. V., Pham, T. T. T., Tran, T. N., & Vo, X. M. (2026). Integrating generative AI and cultural storytelling to enhance geometry learning in Vietnamese primary classrooms: A quasi-experimental study. *Education Sciences*, *16*(4), Article 588. https://doi.org/10.3390/educsci16040588

Norberg, K. A., Almoubayyed, H., De Ley, L., Murphy, A., Weldon, K., & Ritter, S. (2025). Rewriting content with GPT-4 to support emerging readers in adaptive mathematics software. *International Journal of Artificial Intelligence in Education*, *35*(2), 587–626. https://doi.org/10.1007/s40593-024-00420-2

Oğuz Haçat, S., & Kepceoğlu, İ. (2025). Help of ChatGPT about math skills on the subjects of social studies. *The Journal of Educational Research*, *118*(6), 771–782. https://doi.org/10.1080/00220671.2025.2510389

Öner, A. T. (2025). Worked-out examples with ChatGPT-based explanations: A statistics activity. *Journal of Inquiry Based Activities*, *15*, 146–169. https://files.eric.ed.gov/fulltext/EJ1489058.pdf

Oreopoulos, P., Liut, M., Sungu, A., & Low, S. (2026). *Integrating conversational AI into middle-school computer-assisted mathematics learning: A multi-arm randomized trial* (NBER Working Paper No. 35621). National Bureau of Economic Research. https://doi.org/10.3386/w35621

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, *87*(5), 377–385. https://doi.org/10.6028/jres.087.022

Polydoros, G. (2026). *From AI-generated explanations to algebraic reasoning: Teacher-mediated revoicing, justification, and validation in lower secondary algebra* (Version 1, Research Square preprint). https://doi.org/10.21203/rs.3.rs-9806549/v1

Polydoros, G. V. (2025). Harnessing AI and digital tools for micro-credentialing in mathematics education: Enhancing learning outcomes for students with learning difficulties. In *Advances in educational technologies and instructional design* (pp. 195–216). IGI Global. https://doi.org/10.4018/979-8-3693-5488-9.ch009

Prabhakar, A. (2026). Generative AI as a thinking partner: How students' mathematical reasoning develops through dialogue with AI tools. *Journal of Intelligent Decision Making and Information Science*, *3*, Article 2401. https://doi.org/10.17762/jidmis.v3.2401

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E., & Kurd-Misto, E. (2026). *Faster completion, less learning: Generative AI reduced study time on math problems and the knowledge they build* (arXiv:2605.21629). arXiv. https://doi.org/10.48550/arXiv.2605.21629

Rizos, I., Foykas, E., & Georgakopoulos, S. V. (2024). Enhancing mathematics education for students with special educational needs through generative AI: A case study in Greece. *Contemporary Educational Technology*, *16*, Article ep535. https://doi.org/10.30935/cedtech/15487

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher*, *20*(3), 2–9. https://doi.org/10.3102/0013189X020003002

Serin, M. K., & Kepceoğlu, İ. (2026). Generative AI feedback and creative revision in seventh-grade mathematical problem posing. *The Journal of Educational Research*, 1–15. https://doi.org/10.1080/00220671.2026.2735824

So, B., Jung, E., Seong, E., & Kim, H. J. (2026). Mathematical modeling as a discursive process: ChatGPT as a mediating agent in the development of mathematical discourse. *Journal of Educational Research in Mathematics*, *36*(3), 401–431. https://doi.org/10.29275/jerm.2026.36.3.401

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tasdelen, O., & Bodemer, D. (2025). Generative AI in the classroom: Effects of context-personalized learning material and tasks on motivation and performance. *International Journal of Artificial Intelligence in Education*, *35*, 3049–3070. https://doi.org/10.1007/s40593-025-00491-9

Trocado, A., dos Santos, J. M., Saimon, M., & Lavicza, Z. (2026). Learning quadratic functions with ChatGPT: An innovative experience in high school mathematics education. *International Journal of Education in Mathematics, Science and Technology*, *14*(1), 296–314. https://doi.org/10.46328/ijemst.5074

Walkington, C., Pando, M., Lipsmeyer, L. L., Beauchamp, T., Sager, M., & Milton, S. (2025). Middle school girls using generative AI to engage in mathematical problem-posing. *Mathematical Thinking and Learning*. Advance online publication. https://doi.org/10.1080/10986065.2025.2542724

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2025). *Tutor CoPilot: A human–AI approach for scaling real-time expertise* (EdWorkingPaper No. 24-1054 v2 / arXiv:2410.03017v3). Annenberg Institute at Brown University / arXiv. https://doi.org/10.48550/arXiv.2410.03017

Wang, X., & Wei, Y. (2025). The influence of Gen-AI assisted learning on primary school students' math anxiety: An intervention study. *Applied Cognitive Psychology*, *39*(4), Article e70088. https://doi.org/10.1002/acp.70088

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook* (Version 5.0). U.S. Department of Education, Institute of Education Sciences.

Worden, E., Vanacore, K., Haim, A., & Heffernan, N. T. (2025). Scaling effective AI-generated explanations for middle school mathematics in online learning platforms. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25)* (pp. 40–49). ACM. https://doi.org/10.1145/3698205.3729546

Xing, W., Song, Y., Li, C., Liu, Z., Zhu, W., & Oh, H. (2025). Development of a generative AI-powered teachable agent for middle school mathematics learning: A design-based research study. *British Journal of Educational Technology*, *56*(5), 2043–2077. https://doi.org/10.1111/bjet.13586

Xuan, S. H., Nguyen, A. T., Nguyen, T., Nguyen, L., Nguyen, H., Pham, N., Phung, T., Ngo, B., Nguyen, V., Nguyen, M., Tran, T., Le, T., Nguyen, K., & FNU, P. (2025). Evaluating the impact of generative AI in mathematics education: A comparative study in Vietnamese high schools. *Human Behavior and Emerging Technologies*, *2025*(1), Article 8886206. https://doi.org/10.1155/hbe2/8886206

Yunianto, W., Galiç, S., & Lavicza, Z. (2024). Exploring computational thinking in mathematics education: Integrating ChatGPT with GeoGebra for enhanced learning experiences. *International Journal of Education in Mathematics, Science, and Technology*, *12*(6), 1451–1470. https://doi.org/10.46328/ijemst.4437

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CleanednewAI08SupplementS1Q1.pdf