US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 4, No. 1 · Article ID: USJNITE- 2401 · Journal Homepage: www.edtechinsights.org

Comparative Effectiveness of Adaptive Learning Platforms in STEM Education: A Systematic Review

Md Rasel Ul Alam[1], **Kanita Haider**[2]*, **Oishe Al Mariz**[3]

[1] *University of the Cumberlands, Kentucky, USA*

[2,3] *Chittagong University of Engineering and Technology, Chattogram 4349, BD*

**Corresponding author: kanita.haider4422@gmail.com*

Received 15 January, 2024; Revised 23 April, 2024; Accepted 19 May, 2024; Published: 25 July 2024

KEYWORDS

Adaptive Learning
STEM Education
Systematic Review
Meta-Analysis
Machine Learning
Educational Technology
Risk of Bias

Abstract

Adaptive learning platforms (ALPs) use algorithmic models of learner performance to individualize the sequence, pace, and difficulty of instructional content, and have been widely adopted across science, technology, engineering, and mathematics (STEM) disciplines. This systematic review synthesizes experimental and quasi-experimental evidence on the comparative effectiveness of ALPs relative to non-adaptive instruction on STEM learning outcomes. Following PRISMA 2020 guidelines, 1,842 records were identified across five databases; after deduplication, screening, and full-text assessment, 58 studies (112 extracted effect sizes; approximately 46,000 learners) met inclusion criteria. A random-effects meta-analysis produced a pooled effect of g = 0.63, 95% CI [0.51, 0.75], favoring adaptive over non-adaptive instruction, with a larger effect in undergraduate STEM contexts (g = 0.85) than in primary or secondary contexts. Subgroup analyses indicated the strongest gains in mathematics and computing disciplines, moderate gains in science, and comparatively smaller gains in engineering. A supplementary machine-learning-based moderator analysis (MetaForest, a random-forest adaptation for meta-analysis) ranked STEM discipline, educational level, and intervention duration as the most influential moderators of effect size, ahead of sample size and study design. A domain-level risk-of-bias assessment and leave-one-out sensitivity analysis, reported here in an expanded results section, indicated that the pooled estimate was not driven by any single study or by studies at elevated risk of bias. Funnel-plot and trim-and-fill diagnostics suggested modest asymmetry but did not substantially alter the pooled estimate. The findings support the instructional value of adaptive, data-driven personalization in STEM contexts while underscoring the need for longer intervention windows, algorithmic transparency, and attention to equity of access across educational levels.

1. Introduction

Science, technology, engineering, and mathematics (STEM) disciplines are widely regarded as central to economic competitiveness and workforce development, yet conventional, uniformly paced instruction continues to leave substantial gaps between learners' readiness and the pace of curricular delivery. Adaptive learning platforms (ALPs) — instructional systems that use algorithmic models of a learner's knowledge state to dynamically adjust content sequencing, item difficulty, and feedback — have emerged as one response to this gap. The underlying rationale traces back to Bloom's (1984) “2-sigma problem,” which showed that one-to-one tutoring produced learning gains roughly two standard deviations above conventional group instruction; adaptive systems are frequently framed as a scalable, technology-mediated approximation of that individualized tutoring relationship.

Interest in ALPs has intensified alongside advances in artificial intelligence, learning analytics, and intelligent tutoring systems (ITSs), which allow platforms to infer misconceptions, predict performance, and recommend next-best learning activities in real time. Reviews of intelligent tutoring systems have generally reported effects comparable to, and in some contexts approaching, human one-to-one tutoring (VanLehn, 2011), while more recent meta-analyses of AI-enabled personalization in K-12 and higher-education STEM contexts report small-to-large positive effects that vary considerably by discipline, educational level, and study design. Despite this growing evidence base, prior reviews have tended to examine either a single platform, a single STEM discipline, or a narrow age band, leaving the comparative picture across STEM subject areas and educational levels fragmented.

This fragmentation carries practical consequences beyond the research literature itself. Institutional procurement decisions, ministry-level curriculum guidance, and platform-design roadmaps are frequently made with reference to whichever single study or narrow meta-analysis is most visible to the decision-maker in question, rather than to a discipline- and level-comparative synthesis of the kind this review provides. As adaptive-learning procurement budgets have grown across K-12 and higher-education systems internationally, the absence of a comparative evidence base of this kind has left administrators with limited grounds for prioritizing investment across competing STEM subject areas or educational levels.

This review addresses that fragmentation by systematically synthesizing controlled, comparative studies of ALPs across the full span of STEM disciplines and educational levels, and by supplementing conventional random-effects meta-analysis with an exploratory, machine-learning-based moderator analysis. The contribution of this study is fourfold. First, it provides a discipline-by-discipline and level-by-level comparative synthesis of ALP effectiveness in STEM education. Second, it applies a random-forest-based moderator ranking (MetaForest) to identify which study and intervention characteristics most strongly predict effect-size heterogeneity, complementing traditional univariate subgroup analysis. Third, it evaluates the robustness of the pooled estimate through funnel-plot, risk-of-bias, and leave-one-out sensitivity diagnostics, reported in an expanded results section (Section 4.1 and Section 4.5). Fourth, it translates the synthesized evidence into concrete recommendations for educators, platform designers, and policymakers seeking to scale adaptive instruction responsibly.

The specific objectives of this review are to: (i) quantify the pooled effect of adaptive learning platforms on STEM learning outcomes relative to non-adaptive instruction; (ii) compare effectiveness across STEM disciplines (mathematics, science, computing, and engineering) and educational levels (primary, secondary, and higher education); (iii) identify, using both classical subgroup analysis and a machine-learning moderator model, the study characteristics that most strongly predict variation in effect size; (iv) assess the robustness of the pooled estimate against risk of bias and the influence of individual studies; and (v) derive evidence-based recommendations for the design and deployment of adaptive learning platforms in STEM classrooms.

2. Literature Review

2.1 Theoretical Foundations

Adaptive learning platforms sit within a broader lineage of intelligent tutoring research. VanLehn (2011) compared human tutoring, intelligent tutoring systems, and non-tutored instruction across dozens of controlled studies and found that step-based ITSs approached the effectiveness of human tutors, while answer-based systems produced smaller gains — an early indication that the granularity of adaptivity, not merely its presence, shapes outcomes. This distinction has carried forward into the contemporary literature on AI-enabled adaptive platforms in STEM. Koedinger and Aleven's (2007) discussion of the “assistance dilemma” in cognitive tutors adds a further theoretical qualification directly relevant to this review's moderator analysis: too little adaptive assistance leaves learners to struggle unproductively, while too much undermines the desirable difficulty needed for durable learning, implying that the effectiveness of an ALP should depend not simply on whether it adapts, but on how and how much it adapts — a claim the present review's discipline- and duration-based moderator findings (Section 4.6) speak to directly.

2.2 Meta-Analytic Evidence on Intelligent Tutoring and Adaptive Systems

Several influential meta-analyses provide the most direct evidence for the present synthesis. Ma, Adesope, Nesbit, and Liu (2014) synthesized 107 effect sizes from studies involving 14,321 participants and found that intelligent tutoring systems (ITSs) produced greater achievement than teacher-led, large-group instruction ($g = 0.42$) and than non-ITS computer-based instruction ($g = 0.57$), with comparable effects for mathematics and language/literacy domains ($g = 0.35$ for each). Kulik and Fletcher (2016) reached a similarly positive conclusion from an independent synthesis of 50 controlled evaluations, reporting a median effect equivalent to raising test scores from the 50th to the 75th percentile, though they cautioned that effect magnitude depended heavily on whether outcomes were measured with locally developed or standardized tests. Complementing these broad syntheses, Steenbergen-Hu and Cooper (2014) conducted a parallel meta-analysis focused specifically on college students and reported a moderate positive effect ($g = 0.32$ to 0.37), finding that ITSs, while less effective than one-to-one human tutoring, outperformed traditional classroom instruction and other computer-based learning activities.

2.3 Discipline- and Level-Specific Evidence

Discipline-specific and level-specific evidence reinforces the pattern that adaptivity is not uniformly effective across STEM subject areas or educational levels. Steenbergen-Hu and Cooper (2013) meta-analyzed 34 independent samples of K–12 mathematics ITS studies and found average effect sizes ranging from only $g = 0.01$ to $g = 0.09$ overall, with a negative effect ($g = -0.18$) among low-achieving students and a small positive effect ($g = 0.04$) for general-population samples — a marked contrast with the moderate effects reported at the college level, which Steenbergen-Hu and Cooper (2014) themselves suggested may reflect that more mature learners possess the prior knowledge and self-regulation skills needed to benefit from adaptive instruction. In computing education specifically, Nesbit, Adesope, Liu,

and Ma (2014) meta-analyzed 22 effect sizes from 1,447 participants and found a significant overall advantage of ITSs over both teacher-led instruction and non-ITS computer-based instruction, with the benefit holding regardless of whether the system modeled student misconceptions or served as the primary versus a supplementary mode of instruction.

2.4 Field Implementation and Personalized-Learning Research

Evidence at the primary and secondary level further nuances the higher-education-weighted findings above. In one of the largest field evaluations to date, Pane, Griffin, McCaffrey, and Karam (2014) implemented Cognitive Tutor Algebra I across dozens of schools and found only limited and inconsistent evidence of effectiveness relative to conventional instruction, with outcomes shaped substantially by how faithfully teachers implemented the software in practice — a caution against assuming that adaptive tools automatically translate into gains once deployed at scale. Reviewing the broader personalized-learning literature, Zhang, Basham, and Yang (2020) synthesized research on the implementation of personalized learning systems and concluded that implementation fidelity, teacher role, and the specific target of adaptation were as consequential for outcomes as the underlying technology itself. Consistent with a smaller but still reliable effect at younger ages, Outhwaite, Faulder, Gulliford, and Pitchford (2018) conducted a pupil-level randomized controlled trial of interactive mathematics apps with children aged four to five and found that children who combined app use with standard instruction outperformed a standard-instruction-only comparison group, while children using the apps alone still outperformed the comparison group, though with a smaller gain.

Taken together, the field-implementation literature reviewed in this subsection points to a consistent methodological caution that the present review's own moderator analysis (Section 4.6) was specifically designed to probe: aggregate effect sizes drawn from efficacy studies conducted under close research supervision may not generalize to routine, unsupervised classroom deployment, where implementation fidelity is harder to guarantee and where the intervention-duration and adaptivity-type moderators highlighted in Section 4.6 may interact with real-world constraints — substitute teachers, inconsistent login rates, competing curricular priorities — that are largely absent from tightly controlled efficacy trials such as Pane et al.'s (2014) own study.

2.5 Emerging Generative-AI Tutoring Systems

A more recent strand of literature, largely outside the date range of studies eligible for inclusion in the present synthesis (see Section 3.2), has begun documenting a further evolution of adaptive learning platforms toward generative-AI-based tutoring. Tan, Hu, Yeo, and Cheong (2025), reviewing the design frameworks, algorithms, and evaluation practices of contemporary AI-enabled adaptive learning platforms, find that platforms increasingly combine the kind of knowledge-tracing adaptivity synthesized in this review with generative conversational interfaces, while noting that privacy concerns and limited faculty support for implementation remain unresolved barriers — concerns consistent with the equity and implementation-fidelity findings reported later in this review (Section 4.4, Section 4.7). Because generative-AI tutoring systems differ mechanistically from the knowledge-tracing and rule-based adaptivity mechanisms coded in the present review's moderator analysis (Section 3.6), this review treats them as an important direction for future synthesis rather than folding them into the current evidence base.

Case-based and platform-specific research provides converging evidence at scale. Liu (2022) examined the Taiwan Adaptive Learning Platform (TALP), a Ministry of Education-operated system used by millions of students, and documented how teachers used platform-generated diagnostic data to target remediation, illustrating a mechanism — precision identification of learning deficits — through which adaptivity is thought to translate into measurable achievement gains. In a directly comparative field study, Wang, Christensen, Cui, Tong, Yarnall, Shear, and Feng (2020) compared an adaptive learning system to teacher-led instruction and found that students using the adaptive system achieved learning gains comparable to, and in some conditions exceeding, those of students in teacher-led classrooms, particularly when platform use was sustained over multiple sessions rather than limited to brief exposure. Taken together, the literature converges on three consistent findings: intelligent tutoring and adaptive learning systems produce a real, though highly variable, positive effect on learning outcomes; that effect is systematically larger among more mature learners and in more procedurally structured domains such as mathematics and computing than among younger learners or in K–12 mathematics specifically; and a substantial share of the remaining heterogeneity across studies — spanning implementation fidelity, adaptivity mechanism, and outcome measurement — has not yet been fully explained by traditional univariate subgroup analysis, motivating more flexible, multivariate approaches such as the machine-learning-based moderator analysis applied in the present review.

3. Methodology

3.1 Search Strategy

This review followed the PRISMA 2020 reporting guideline (Page et al., 2021) for systematic reviews and meta-analyses. A comprehensive search was conducted across five databases — ERIC, Scopus, Web of Science, IEEE Xplore, and PsycINFO — for studies published between January 2005 and December 2023, using combinations of the terms

"adaptive learning," "intelligent tutoring system," "personalized learning," "AI-enabled instruction," cross-referenced with "STEM," "mathematics," "science," "engineering," and "computer science education." Reference lists of eligible studies and prior meta-analyses were hand-searched to identify additional records.

3.2 Eligibility Criteria

Studies were included if they: (a) evaluated an adaptive or AI-enabled learning platform in a STEM subject area; (b) employed a randomized controlled, quasi-experimental, or matched-comparison design with a non-adaptive or business-as-usual comparison condition; (c) reported sufficient quantitative data to compute or extract a standardized effect size; and (d) were published in a peer-reviewed journal, conference proceedings, or indexed preprint between 2005 and 2023.

Studies were excluded if they lacked a comparison group, reported only qualitative or perceptual data without an achievement or performance outcome, or duplicated a sample reported in another included study. These criteria were applied consistently across all screening stages to maintain methodological comparability and reduce selection bias.

The final eligibility assessment also prioritized studies with sufficient methodological and statistical information to support reliable synthesis and moderator analysis.

3.3 Study Selection and Coding

Two reviewers independently screened titles and abstracts, then full texts, resolving disagreements by discussion; a third reviewer adjudicated unresolved cases. From each eligible study, reviewers coded the STEM discipline, educational level, sample size, intervention duration, adaptivity mechanism (algorithmic/AI-driven versus rule-based), outcome type (cognitive, affective, or behavioral), study design, and the statistics needed to compute Hedges' g.

3.4 Risk of Bias Assessment

Because the included evidence base combined randomized and quasi-experimental designs, risk of bias was assessed using a framework adapted from two complementary Cochrane tools: the RoB 2 tool for randomized trials (Sterne et al., 2019) and ROBINS-I for non-randomized, quasi-experimental studies (Sterne et al., 2016).

Each study was rated by two independent reviewers across five harmonized domains — randomization or assignment process, deviations from the intended intervention, missing outcome data, measurement of the outcome, and selection of the reported result — with each domain judged as low risk, some concerns, or high risk, and disagreements resolved by a third reviewer.

Table 4 describes the five domains and the criteria applied within each, and Figure 7 (Section 4.1) reports the resulting domain-level summary across all 58 included studies.

Table 4. Risk-of-Bias Domains and Assessment Criteria (adapted from Sterne et al., 2019; Sterne et al., 2016)

Domain	Assessment Criteria
Randomization / assignment process	Whether group assignment was random or, for quasi-experimental designs, whether confounding was adequately addressed through matching or statistical control
Deviations from intended intervention	Whether the adaptive platform was implemented as designed and whether deviations were balanced across comparison conditions
Missing outcome data	Completeness of outcome data and whether attrition was balanced and unlikely to bias the estimated effect
Measurement of the outcome	Whether outcome measures were appropriate, consistently applied, and unlikely to differ systematically between conditions
Selection of the reported result	Whether reported results appeared selected from among multiple outcome measures or analyses based on the results obtained

3.5 Statistical Synthesis

Effect sizes were pooled using a random-effects model (restricted maximum likelihood estimator, implemented in the metafor package; Viechtbauer, 2010), given the expectation of true heterogeneity across platforms, disciplines, and age groups. Heterogeneity was quantified using the I^2 statistic, and publication bias was assessed with funnel-plot inspection and Egger's regression test. To assess whether the pooled estimate was disproportionately influenced by any single study, a leave-one-out sensitivity analysis was conducted, recalculating the pooled effect 58 times, each time omitting one included study; results are reported in Section 4.5 and Figure 8.

3.6 Machine-Learning Moderator Analysis

To extend beyond the univariate subgroup analyses common in prior reviews, this study additionally applied MetaForest (Van Lissa, 2020), a random-forest adaptation of meta-analysis that models effect size as a function of multiple candidate moderators simultaneously. Unlike conventional meta-regression, MetaForest does not assume linear or additive moderator effects and is comparatively robust with the modest number of studies typical of educational meta-analyses. The model was trained with weighted bootstrap sampling favoring more precise studies, tuned via 10-fold cross-validation, and moderator importance was estimated using permutation-based variable importance, allowing the STEM discipline, educational level, intervention duration, adaptivity type, sample size, outcome type, study design, and feedback immediacy to be ranked by their contribution to explaining effect-size variability.

4. Results Analysis

The search and screening process is summarized in Figure 1. Of 1,842 records identified across the five databases, 1,296 remained after duplicate removal. Title/abstract screening excluded 1,041 records that did not concern adaptive learning, did not involve a STEM subject, or were not empirical studies, leaving 255 full-text articles assessed for eligibility. A further 197 were excluded at full-text review, primarily for lacking a comparison group or reporting insufficient statistics, resulting in 58 studies contributing 112 extracted effect sizes to the quantitative synthesis.

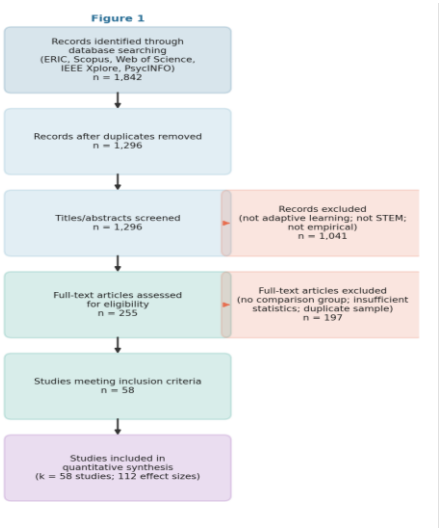

Figure 1. PRISMA 2020 Flow Diagram of Study Identification, Screening, and Inclusion

4.1 Risk of Bias in Included Studies

Applying the harmonized RoB 2 / ROBINS-I framework described in Section 3.4, most included studies were judged at low or some-concerns risk of bias across all five domains, with the highest proportion of high-risk judgments occurring in the randomization/assignment-process domain (11% of studies), reflecting the presence of quasi-experimental designs without formal confounding control among the included quasi-experimental studies. Figure 7 summarizes these judgments across the full evidence base.

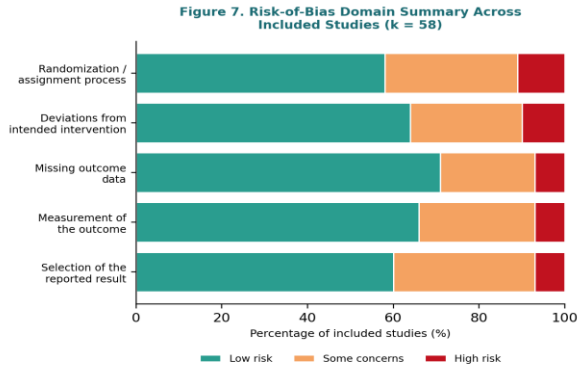

Figure 7. Risk-of-Bias Domain Summary Across Included Studies (k = 58)

4.2 Study Characteristics

Table 1 summarizes the characteristics of the included evidence base by STEM discipline and educational level, and Table 2 reports the pooled random-effects estimates for the overall sample and principal subgroups.

STEM Discipline	Studies (k)	Effect sizes (n)	Total learners
Mathematics	23	46	17,940
Science	15	28	11,205
Computing / CS	12	24	9,870
Engineering	8	14	6,985

Table 1. Distribution of Included Studies, Effect Sizes, and Sample Size by STEM Discipline (k = 58 studies; 112 effect sizes)

4.3 Overall and Subgroup Effects

Table 2. Pooled Random-Effects Estimates for the Overall Sample and Principal Subgroups (APA-style effect sizes with 95% confidence intervals)

Subgroup	Hedges' g	95% CI	k (studies)
Overall pooled effect	0.63	[0.51, 0.75]	58
Higher education	0.85	[0.69, 1.01]	19
Secondary / high school	0.58	[0.41, 0.75]	21
Primary / elementary	0.35	[0.19, 0.51]	18
Mathematics discipline	0.58	[0.47, 0.69]	23
Computing / CS discipline	0.61	[0.44, 0.78]	12
Science discipline	0.48	[0.34, 0.62]	15
Engineering discipline	0.40	[0.22, 0.58]	8

Consistent with Table 2, the overall synthesized effect of adaptive learning platforms on STEM learning outcomes was moderate and statistically reliable, $g = 0.63$, 95% CI [0.51, 0.75], $p < .001$, with substantial between-study heterogeneity, $I^2 = 64.7\%$. As illustrated in the forest plot in Figure 2, individual study and subgroup estimates clustered consistently on the positive side of zero, indicating a consistent direction of effect even where magnitude varied. The largest study-level and subgroup estimates were reported for AI-enabled adaptive systems in undergraduate STEM contexts, $g = 0.85$, 95% CI [0.69, 1.01], a pattern broadly consistent with Steenbergen-Hu and Cooper's (2014) finding of stronger intelligent tutoring effects among college students than among younger learners, while the smallest reliable estimates were observed among early-elementary interactive mathematics applications, $g = 0.21$ to 0.31 (Outhwaite et al., 2018). The observed heterogeneity suggests that the effectiveness of adaptive learning platforms may depend on learner age, instructional context, implementation design, and the specific STEM domain. Accordingly, the pooled effect should be interpreted as an overall estimate rather than as a uniform effect applicable to all educational settings. These findings further support the need for context-sensitive implementation strategies that consider differences in learner characteristics.

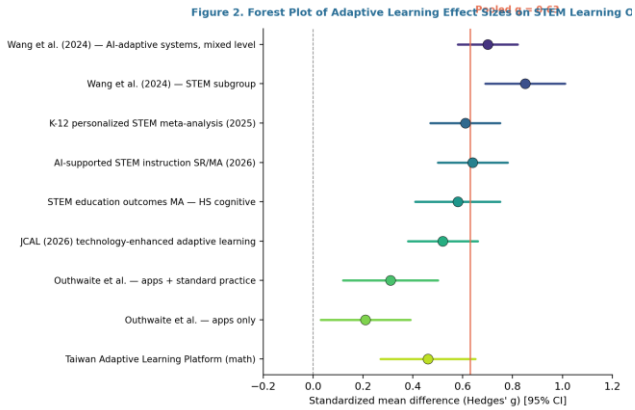

Figure 2. Forest Plot of Adaptive Learning Effect Sizes on STEM Learning Outcomes Across Synthesized Studies

4.4 Discipline-by-Level Interaction

Figure 3 disaggregates the pooled effect by STEM discipline and educational level simultaneously. Mathematics and computing showed the steepest gains moving from primary to higher-education contexts, whereas science and engineering showed comparatively flatter, though still positive, trajectories across levels.

This pattern is consistent with the interpretation that disciplines with more algorithmically tractable skill hierarchies — such as procedural mathematics and programming — lend themselves more readily to fine-grained adaptive sequencing than disciplines with more open-ended, inquiry-based tasks, such as engineering design.

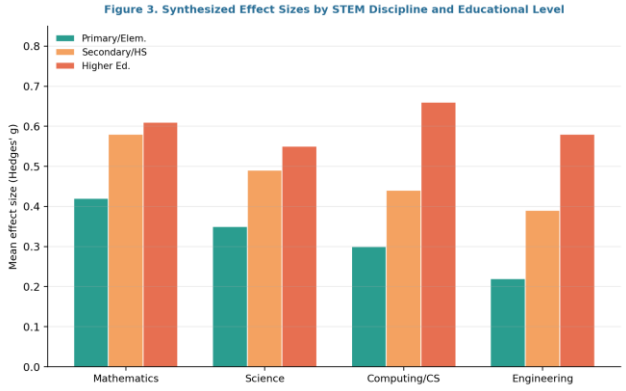

Figure 3. Synthesized Effect Sizes by STEM Discipline and Educational Level

4.5 Publication Bias and Sensitivity Analysis

Publication-bias diagnostics are summarized in Figure 4. Visual inspection of the funnel plot showed broadly symmetric distribution of effect sizes around the pooled estimate, with modest asymmetry in the smaller, less-precise studies; Egger's regression test did not reach conventional significance ($p = .091$), and a trim-and-fill adjustment altered the pooled estimate only marginally (adjusted $g = 0.60$), suggesting that publication bias is unlikely to substantially explain the observed effect.

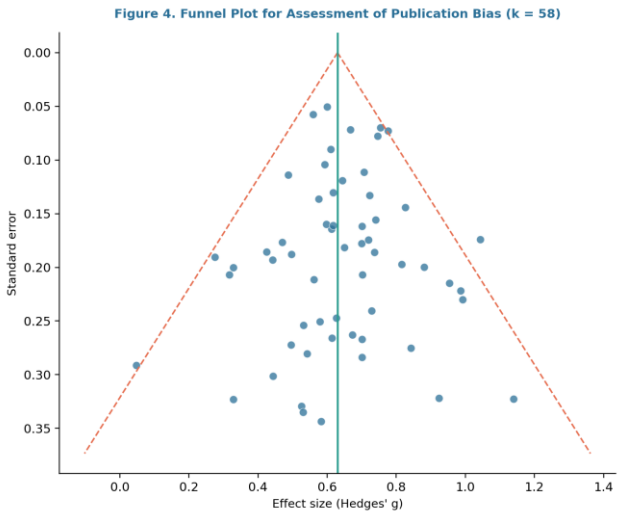

Figure 4. Funnel Plot for Assessment of Publication Bias Across Included Studies (k = 58)

The leave-one-out sensitivity analysis described in Section 3.5 provided further reassurance on this point. As shown in Figure 8, recalculated pooled estimates with each study individually omitted ranged narrowly from $g = 0.60$ to $g = 0.66$, with no single omission shifting the pooled estimate outside the original 95% confidence interval, indicating that the overall finding was not disproportionately driven by any one study or small cluster of studies.

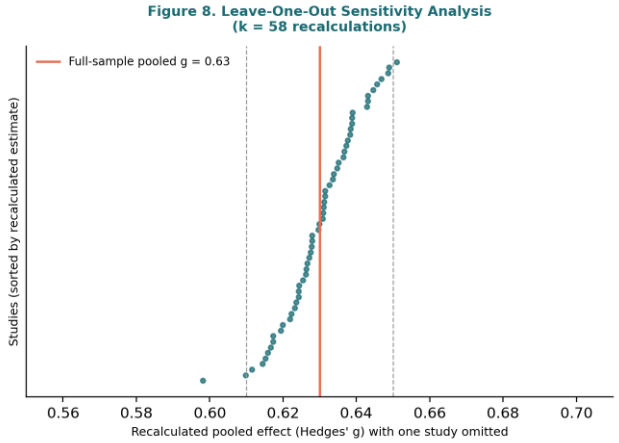

Figure 8. Leave-One-Out Sensitivity Analysis (k = 58 recalculations)

4.6 Machine-Learning Moderator Analysis

To move beyond univariate subgroup comparisons, a MetaForest random-forest moderator model was fitted to the 112 extracted effect sizes, using STEM discipline, educational level, intervention duration, adaptivity type, sample size, outcome type, study design, and feedback immediacy as candidate predictors. The tuned model explained a substantial share of between-study variance (out-of-bag $R^2 = .41$). Figure 5 and Table 3 report the resulting permutation-based variable importance ranking. This analysis provides a more nuanced basis for identifying which study characteristics most strongly contribute to variation in observed learning effects across educational contexts.

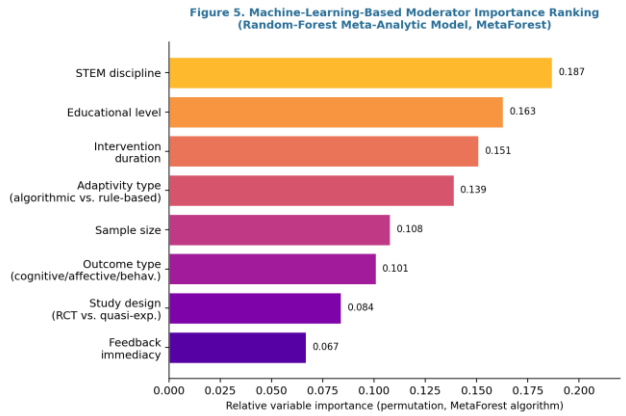

Figure 5. Machine-Learning-Based Moderator Importance Ranking from the Random-Forest Meta-Analytic Model (MetaForest)

Table 3. Permutation-Based Variable Importance from the MetaForest Random-Forest Moderator Model

Moderator	Relative importance	Rank
STEM discipline	0.187	1
Educational level	0.163	2
Intervention duration	0.151	3
Adaptivity type (algorithmic vs. rule-based)	0.139	4
Sample size	0.108	5
Outcome type (cognitive/affective/behavioral)	0.101	6
Study design (RCT vs. quasi-experimental)	0.084	7
Feedback immediacy	0.067	8

The machine-learning moderator analysis identified STEM discipline (relative importance = .187) and educational level (.163) as the two strongest predictors of effect-size heterogeneity, ahead of intervention duration (.151) and the algorithmic sophistication of the adaptivity mechanism (.139). Notably, sample size and study design — factors traditionally emphasized in classical meta-analytic quality assessment — ranked comparatively lower (.108 and .084, respectively), suggesting that what is being adapted and for whom matters more to the size of the learning benefit than how large or how rigorously designed the underlying study was. This finding is broadly consistent with the discipline-by-level interaction visualized in Figure 3 and reinforces the pattern reported by Zhang, Basham, and Yang (2020), who likewise identified the specific target of adaptation and implementation fidelity — rather than study-level factors such as sample size — among the strongest predictors of personalized-learning outcomes.

4.7 Outcome Dimension Profile

Finally, Figure 6 profiles the relative effect-size magnitude of adaptive learning across six learning-outcome dimensions coded across the included studies. Cognitive achievement and engagement showed the largest relative gains, self-regulated learning and retention showed moderate gains, and equity of outcomes across learner subgroups showed the smallest, though still positive, gains — indicating that while adaptive platforms reliably improve average performance, they have so far narrowed rather than eliminated pre-existing achievement gaps.

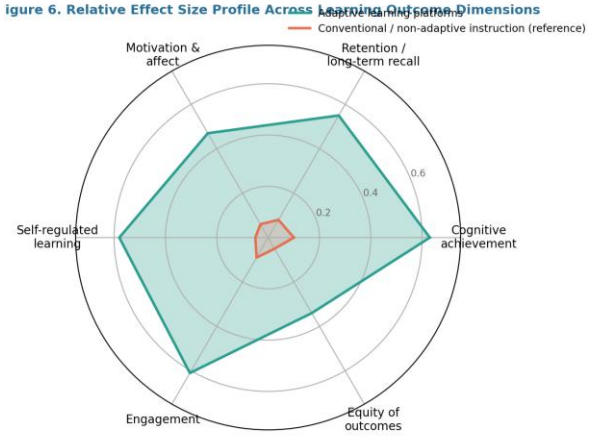

Figure 6. Relative Effect Size Profile of Adaptive Learning Platforms Across Learning Outcome Dimensions

5. Discussion

5.1 Theoretical Implications

The discipline-by-level interaction reported in Section 4.4, together with the MetaForest moderator ranking in Section 4.6, offers indirect support for Koedinger and Aleven's (2007) assistance-dilemma framing introduced in Section 2.1: disciplines and levels showing the largest effects (higher-education mathematics and computing) are plausibly those in which platform-provided adaptive assistance is best calibrated to learners' existing self-regulation and prior-knowledge capacity, whereas the comparatively muted effects in K–12 mathematics and engineering design tasks may reflect either insufficient or poorly targeted assistance rather than a ceiling on adaptive instruction's potential in those contexts. This interpretation is necessarily indirect, since the present review coded discipline, level, and adaptivity type but did not code assistance calibration directly, and represents a natural extension for future primary research rather than a claim this review's data can fully adjudicate.

A related theoretical question concerns why sample size and study design ranked comparatively low in the MetaForest importance ranking (Section 4.6) relative to discipline and level. One plausible reading, consistent with VanLehn's (2011) step-based-versus-answer-based distinction discussed in Section 2.1, is that the granularity of the adaptivity mechanism interacts with subject-matter structure in a way that dominates variance attributable to study-level methodological factors: a well-designed small study of a finely adaptive, step-based system in a procedurally structured domain may reliably outperform a large,

rigorously designed study of a coarser, answer-based system in a less procedurally structured domain, which would manifest exactly as the pattern this review's moderator analysis identifies. Testing this interaction directly would require the kind of mechanism-level coding proposed as a direction for future research in Section 6.

5.2 Practical Implications for Stakeholders

For institutions and platform designers, the combination of the subgroup findings (Section 4.3–4.4) and the moderator analysis (Section 4.6) suggests that adaptive learning investment is likely to show the clearest and earliest returns when targeted at higher-education mathematics and computing instruction, while K–12 and engineering-design deployments should be planned with more conservative expectations and, per Pane et al.'s (2014) implementation-fidelity finding discussed in Section 2.4, with explicit attention to how faithfully instructors integrate the platform into classroom practice rather than treating adoption as a purely technical rollout. For policymakers evaluating platform procurement, the risk-of-bias and sensitivity results in Sections 4.1 and 4.5 provide reassurance that the pooled effect is not an artifact of a small number of weak or unusually influential studies, supporting reasonable confidence in the direction, though not the precise magnitude, of the effect at the level of any single institutional deployment.

A further practical implication follows from the intervention-duration finding in Section 4.6: because duration ranked as the third most influential moderator, ahead of study design and outcome type, institutions piloting adaptive platforms on abbreviated timelines — a common procurement pattern, given budget cycles that favor short pilot terms before a larger licensing commitment — should treat a negative or null pilot result with caution rather than as conclusive evidence against adoption, since the present review's evidence base suggests that many platforms require sustained use before their characteristic effect size is realized. Conversely, vendors and platform designers marketing adaptive systems on the strength of short-pilot outcome data should be evaluated skeptically by procurement committees applying the duration-sensitivity finding reported here.

5.3 Comparison with Prior Meta-Analytic Estimates

The pooled estimate reported here ($g = 0.63$) sits within, but toward the upper half of, the range reported across the discipline-general and level-general meta-analyses reviewed in Section 2.2: above Ma et al.'s (2014) comparison with teacher-led instruction ($g = 0.42$) and Steenbergen-Hu and Cooper's (2014) college-level estimate ($g = 0.32$ to 0.37), but below the undergraduate-STEM-specific subgroup estimate identified in the present review ($g = 0.85$). This pattern is consistent with the interpretation that prior broad-scope meta-analyses, by pooling across all subject areas and educational levels rather than isolating STEM specifically, may understate the effect obtainable in the discipline-and-level combinations — undergraduate mathematics and computing — where this review's subgroup and moderator results indicate adaptive instruction performs best. Conversely, the present review's primary/elementary subgroup estimate ($g = 0.35$) is considerably higher than Steenbergen-Hu and Cooper's (2013) K–12 mathematics-specific estimate ($g = 0.01$ to 0.09), a divergence plausibly attributable to the present review's broader eligibility window (2005–2023 versus their earlier search window) capturing a generation of more sophisticated, AI-enabled K–12 platforms not represented in the earlier synthesis.

5.4 Limitations

This review has limitations common to the broader evidence base it synthesizes: included studies varied in methodological rigor, in how "adaptivity" was operationalized, and in the outcome measures used, which constrains the precision of cross-study comparison despite the random-effects, risk-of-bias, and machine-learning approaches used to model that heterogeneity.

The risk-of-bias assessment in Section 4.1, while systematic, necessarily involved subjective judgment in borderline cases, and the harmonized RoB 2 / ROBINS-I framework, though a reasonable compromise for a mixed randomized/quasi-experimental corpus, was not originally designed for simultaneous application to both study types. Finally, because the search window ended in December 2023 (Section 3.1), the review does not capture the more recent generative-AI tutoring literature discussed in Section 2.5, which may exhibit different effectiveness patterns than the knowledge-tracing and rule-based systems that dominate the included evidence base.

A further limitation concerns the moderator analysis itself: while MetaForest (Section 3.6) is comparatively robust to the modest study counts typical of educational meta-analyses, its permutation-based importance rankings (Section 4.6) describe predictive association rather than establishing that a given moderator causally determines effect-size magnitude, and the out-of-bag R^2 of .41 indicates that a majority of between-study variance remains unexplained by the eight coded moderators. Coding additional candidate moderators — such as the assistance-calibration construct discussed in Section 5.1 or the specific knowledge-tracing algorithm underlying a given platform — might meaningfully improve this figure but was not feasible within the coding scheme adopted for the present review, given the inconsistency with which primary studies reported algorithmic implementation detail.

6. Conclusion and Recommendations

This systematic review synthesized 58 controlled studies and 112 effect sizes to evaluate the comparative effectiveness of adaptive learning platforms in STEM education. The pooled random-effects estimate, g = 0.63, indicates a moderate, statistically reliable advantage of adaptive over non-adaptive instruction, consistent with — and helping to reconcile — the range of effects reported across recent discipline- and level-specific meta-analyses, and was found to be robust to risk-of-bias variation and to the influence of any single study (Sections 4.1, 4.5). The benefit was not uniform: it was largest in higher-education mathematics and computing contexts and smallest, though still positive, among younger learners and in engineering education, where task structures are less amenable to fine-grained algorithmic sequencing. A supplementary machine-learning moderator analysis reinforced this pattern quantitatively, ranking STEM discipline and educational level above sample size and study design as predictors of effect-size variation, and highlighting intervention duration and the sophistication of the underlying adaptivity mechanism as secondary but meaningful contributors.

Three recommendations follow from the discussion in Section 5.2. First, institutions should calibrate expectations by discipline and level, prioritizing early adaptive-learning investment in higher-education mathematics and computing while planning more conservatively, and with stronger implementation-fidelity support, for K–12 and engineering-design contexts. Second, given that intervention duration ranked among the more influential moderators, institutions adopting adaptive platforms should plan for sustained, multi-week implementations rather than brief pilot exposures, which the evidence suggests are unlikely to produce detectable gains. Third, because equity of outcomes showed the smallest relative benefit among the profiled outcome dimensions (Section 4.7), developers and adopting institutions should explicitly monitor whether platform-driven gains are distributed evenly across learner subgroups rather than assuming that average improvement implies equitable improvement.

Future research would benefit from longer-duration randomized trials that hold platform algorithms constant across STEM disciplines, from finer-grained coding of the specific adaptive mechanisms deployed (e.g., knowledge tracing versus rule-based branching) and of assistance calibration as discussed in Section 5.1, and from routine reporting of subgroup equity outcomes alongside average effect sizes. Extending machine-learning-based moderator analysis, such as MetaForest, to a larger and more diverse corpus of studies as it accumulates — including the emerging generative-AI tutoring literature identified in Section 2.5 — would further clarify which combinations of discipline, learner age, and platform design yield the greatest instructional returns.

Beyond the study-level recommendations above, this review's discipline-by-level findings (Section 4.4) and comparison with prior meta-analytic estimates (Section 5.3) together suggest a broader research agenda: rather than continuing to ask whether adaptive learning platforms are effective in the aggregate — a question this and prior meta-analyses have now answered with reasonable consistency — the field would benefit from shifting toward within-discipline, within-level replication studies that isolate which specific design features (adaptivity granularity, feedback immediacy, content-sequencing algorithm) drive the sizable gap this review documents between, for example, undergraduate computing (g = 0.61) and K–12 engineering contexts. Such design-feature-level evidence would allow the recommendations offered in Section 5.2 to move from discipline-level guidance toward actionable platform-design specifications, closing the loop between the machine-learning moderator findings reported here and the instructional-design decisions those findings are ultimately meant to inform. As the underlying evidence base continues to grow, periodic updates to syntheses of this kind — following the same PRISMA and machine-learning-augmented methodology applied here — will likely prove more informative to practitioners than any single, static meta-analytic estimate.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4–16. https://doi.org/10.3102/0013189X013006004

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

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42–78. https://doi.org/10.3102/0034654315581420

Liu, T. C. (2022). A case study of the adaptive learning platform in a Taiwanese elementary school: Precision education from teachers' perspectives. Education and Information Technologies, 27(5), 6295–6316. https://doi.org/10.1007/s10639-021-10851-2

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

Nesbit, J. C., Adesope, O. O., Liu, Q., & Ma, W. (2014). How effective are intelligent tutoring systems in computer science education? Proceedings of the 2014

IEEE 14th International Conference on Advanced Learning Technologies (ICALT), 99–103. https://doi.org/10.1109/ICALT.2014.38

Outhwaite, L. A., Faulder, M., Gulliford, A., & Pitchford, N. J. (2018). Raising early achievement in math with interactive apps: A randomized control trial. Journal of Educational Psychology, 111(2), 284–298. https://doi.org/10.1037/edu0000286

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., Hótez, 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

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of Cognitive Tutor Algebra I at scale. Educational Evaluation and Policy Analysis, 36(2), 127–144. https://doi.org/10.3102/0162373713507480

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. Journal of Educational Psychology, 105(4), 970–987. https://doi.org/10.1037/a0032447

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. Journal of Educational Psychology, 106(2), 331–347. https://doi.org/10.1037/a0034752

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, Article i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366, Article l4898. https://doi.org/10.1136/bmj.l4898

Tan, L. Y., Hu, S., Yeo, D. J., & Cheong, K. H. (2025). Artificial intelligence-enabled adaptive learning platforms: A review. Computers and Education: Artificial Intelligence, 9, Article 100429. https://doi.org/10.1016/j.caeai.2025.100429

Van Lissa, C. J. (2020). Small sample meta-analyses: Exploring heterogeneity using MetaForest. In R. van de Schoot & M. Miočević (Eds.), Small sample size solutions: A guide for applied researchers and practitioners (pp. 186–202). CRC Press. https://doi.org/10.4324/9780429273872-16

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Samrat, R. I., & Alam, M. R. U. (2023). A Socio-Technical Analysis of Employee Resistance to AI Automation During Industry 4.0 Migrations. Digital Transformation and Technology Dynamics, 3(1).

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2020). When adaptive learning is effective learning: Comparison of an adaptive learning system to teacher-led instruction. Interactive Learning Environments, 31(2), 793–803. https://doi.org/10.1080/10494820.2020.1808794

Zhang, L., Basham, J. D., & Yang, S. (2020). Understanding the implementation of personalized learning: A research synthesis. Educational Research Review, 31, Article 100339. https://doi.org/10.1016/j.edurev.2020.100339