Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1646

The Effects of Standalone Digital Learning Programs on PreK–12 Mathematics and Science Achievement: A Meta-Analysis of Randomized Controlled Trials

Qi Zhang
American Institutes for Research

Laura Michaelson
Denver Public Schools

Kamal Middlebrook
American Institutes for Research

Sarah Sahni
Abdul Latif Jamell Poverty Action Lab

Natalie Bills
American Institutes for Research

Most reviews of educational technology pool programs that supplement teachers with programs that teach on their own. We synthesized 68 randomized controlled trials (147 effect sizes) of standalone digital learning programs in PreK–12 mathematics and science, compared with traditional in-person classroom instruction. Using correlated and hierarchical effects models with robust variance estimation, we tested moderators in blocks, adjusting for study characteristics. Students who used standalone programs outperformed comparison students by an average of g = 0.47 (95% CI [0.35, 0.60]), an estimate that held across sensitivity analyses with no evidence of selective reporting. Effects varied widely, with programs that gave personalized feedback or paired practice with direct instruction producing larger effects than programs without personalization or with practice alone. Effects were also larger in low- and middle-income countries than in high-income countries, and larger in studies with a conflict of interest. Because most programs were used at school for part of students' instructional time, the findings describe what well-designed programs can contribute within regular schooling.

VERSION: November 2026

Suggested citation: Zhang, Qi, Laura Michaelson, Kamal Middlebrook, Sarah Sahni, and Natalie Bills. (2026). The Effects of Standalone Digital Learning Programs on PreK–12 Mathematics and Science Achievement: A Meta-Analysis of Randomized Controlled Trials. (EdWorkingPaper: 26-1646). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/x38n-mh30nstitute at Brown University: https://doi.org/10.26300/4a9m-1f54

The Effects of Standalone Digital Learning Programs on PreK–12 Mathematics and Science Achievement: A Meta-Analysis of Randomized Controlled Trials

Qi Zhang[1*]

Laura Michaelson[2]

Kamal Middlebrook[1]

Sarah Sahni[3]

Natalie Bills[1]

[1]American Institutes for Research (AIR), 1400 Crystal Drive, 10th Floor, Arlington, VA 22202, USA

[2]Denver Public Schools, 1860 Lincoln Street Floor 11, Denver, CO 80203, USA

[3]Abdul Latif Jamell Poverty Action Lab (J-PAL), 400 Main Street, E19-201, Cambridge, MA 02142, USA

[*]Corresponding author. Email: qzhang@air.org

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant No. 2200883. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Abstract

Most reviews of educational technology pool programs that supplement teachers with programs that teach on their own. We synthesized 68 randomized controlled trials (147 effect sizes) of standalone digital learning programs in PreK–12 mathematics and science, compared with traditional in-person classroom instruction. Using correlated and hierarchical effects models with robust variance estimation, we tested moderators in blocks, adjusting for study characteristics. Students who used standalone programs outperformed comparison students by an average of g = 0.47 (95% CI [0.35, 0.60]), an estimate that held across sensitivity analyses with no evidence of selective reporting. Effects varied widely, with programs that gave personalized feedback or paired practice with direct instruction producing larger effects than programs without personalization or with practice alone. Effects were also larger in low- and middle-income countries than in high-income countries, and larger in studies with a conflict of interest. Because most programs were used at school for part of students' instructional time, the findings describe what well-designed programs can contribute within regular schooling.

Introduction

Mathematics and science achievement in the PreK–12 years shapes how far students go in school, which careers are open to them, and the nation's capacity for scientific and technical work. Yet recent national trends in both subjects raise concerns about whether students are developing these foundational skills. On the National Assessment of Educational Progress (NAEP) long-term trend assessment, 9-year-olds scored lower in mathematics in 2022 than in 2020, and the drop was largest among lower-performing students (National Center for Education Statistics [NCES], 2026). Similar concerns are evident in science. Fourth graders scored lower on the 2019 NAEP science assessment than they did in 2015 (NCES, 2021).

Improving mathematics and science learning will require schools to make informed choices about the instructional tools they use, including a rapidly growing range of digital learning products. Many districts have expanded their use of courseware, tutoring systems, simulations, and practice apps, particularly as schools recover from the pandemic. These are large investments with uncertain returns. Digital programs can change how students spend instructional time, stretch limited staff, and make practice more engaging. They can also widen opportunity gaps when students differ in their access to devices, reliable internet, and adults who can help them use the programs (Irwin et al., 2023). Schools need more than an answer to whether educational technology works. They need to know which kinds of programs improve learning, for which students, and under what conditions.

Defining Digital Instruction

Answering these questions requires being clear about what we mean by digital instruction We use the term to describe a broad range of ways students learn through technology, from fully online schools and blended models that combine teacher-led and online instruction to student-facing programs that provide targeted lessons or practice. These approaches differ considerably

in how they are designed, implemented, and used with students, so we do not treat digital instruction as a single type of intervention (Means et al., 2013; Means et al., 2014).

This paper focuses on one part of this broader landscape, which we call standalone digital learning programs. These are student-facing technologies that deliver mathematics or science content for a portion of students' instructional time, so that students can learn from them without a teacher presenting the material. Examples include courseware modules, tutoring systems, interactive practice programs, and simulations that give students explanations, guided practice, feedback, or content matched to their skill level. Many of these programs are used during the school day, including in a classroom or computer lab with a teacher present. What makes them standalone is that the teacher's role during use is limited to helping students get started and stay on task. This distinction lets us examine what happens when students receive mathematics or science instruction directly through a digital program. Digital tools that supplement content delivered primarily by a teacher fall outside this focus.

Mixed Findings in Prior Work

Prior research reinforces the importance of distinguishing among different forms of digital instruction. At one end of the spectrum, fully online schools and courses, where students receive most or all of their instruction online, have generally produced less favorable results than in-person instruction. In Ohio, students who attended school mostly or entirely online scored lower on state tests than similar students in traditional schools (Ahn & McEachin, 2017). In another study linking about seven million online course records to student outcomes, taking courses online was associated with lower math and reading scores (Heinrich et al., 2018). Because students choose whether to enroll in online schools and courses, however, these studies

cannot fully separate the effects of online instruction from differences between students who do and do not choose these options.

At the other end of the spectrum, results have been more favorable for digital technologies used to support teacher-led instruction. Evaluations of programs that provide guided practice, feedback, and interactive visuals have found gains in mathematics and science achievement compared with regular instruction alone (Finkelstein et al., 2005; Ritter et al., 2007; Roschelle et al., 2010). These findings suggest that digital programs can support student learning, but they tell us less about what happens when a program itself delivers the instructional content. That conclusion is hard to test, however, because most of these programs were never designed or evaluated as a replacement for teacher instruction. This leaves an important question about the programs at the center of this paper: How effective are standalone digital learning programs when they deliver instruction for part of students' learning time?

Broader reviews point to the same need for a more focused synthesis. A meta-analysis comparing online and blended learning with face-to-face instruction found more consistent benefits for blended designs than for fully online ones. The same review found that the benefits often came with extra learning time and resources, not from the online format itself (Means et al., 2013). Reviews of computer-assisted instruction and related classroom technologies have also reported positive average effects in mathematics and science, alongside substantial variation across programs, settings, and outcomes (Cheung & Slavin, 2013; Cheung et al., 2017). An average effect can therefore conceal meaningful differences in how programs are designed, how they are used, and the students and settings in which they are implemented (Means et al., 2014). For schools, this variation makes it difficult to know which programs are likely to improve learning. For researchers, it points to the need for a synthesis that goes beyond the average effect

to examine which program features and study conditions are associated with differences in outcomes.

Pedagogical Mechanisms

The variation in effects across digital learning programs may reflect differences in the instruction they provide. We therefore think about standalone programs in terms of the instruction they provide, not the technology they use. This view comes from several lines of research on how students learn, and each one points to program features we test as moderators.

The first line of research concerns tutoring. One-to-one tutoring is among the most effective forms of instruction (VanLehn, 2011), and much of its advantage comes from two things a tutor does well. A tutor responds to each answer with feedback, and adjusts what comes next to what the student knows. Computer tutors that copy these moves can produce gains close to those of human tutors (Kulik & Fletcher, 2016; VanLehn, 2011). Research on feedback further suggests that its effectiveness depends on the information students receive about their performance and how they can improve (Hattie & Timperley, 2007; Wisniewski et al., 2020). We therefore expect programs that give feedback, and programs that adapt to a student's level, to produce larger gains than programs that do neither. Our personalization moderator captures these features.

The second line comes from research on cognitive load and guided instruction. Students who are new to a topic learn more from explicit explanations and worked examples than from practice or problem solving alone, because novices have little prior knowledge to organize new information (Kirschner et al., 2006). This may be particularly important for standalone programs because the program itself is responsible for delivering the instructional content. A program that pairs practice with direct instruction can show students why a procedure works. A practice-only

program leaves that work to the student. We therefore expect programs that combine practice with direct instruction to produce larger effects than programs that rely on practice alone.

The third line concerns discovery and collaborative learning. Simulations, virtual labs, and other exploratory activities let students manipulate abstract ideas and test their own predictions. That kind of activity can build conceptual understanding in mathematics and science. Research suggests that discovery learning is more effective when students receive guidance than when they are left to discover concepts on their own (Alfieri et al., 2011; Mayer, 2004). This suggests that discovery and collaborative activities may be especially useful when combined with other forms of instructional support. Our instructional practice moderator tests these combinations, using practice alone as the reference group and comparing programs that add direct instruction, collaboration or discovery, or both.

The Importance of Self-Directed Learning

Standalone digital learning programs also change who is responsible for managing the learning process. When a program takes over part of what a teacher does, students take on more of the work of managing their own learning. They have to stay focused, keep going when a problem is hard, notice when they do not understand, and decide how to respond. Theories of self-regulated learning describe these skills, which develop over time and vary across students (Zimmerman, 2002). Research on computer-based learning also suggests that students' ability to regulate their learning can influence how successfully they learn in these environments (Azevedo, 2005). This view points to several conditions that may shape the effectiveness of standalone programs, including the support students receive, how much they use the program, and the context in which they use it.

One is support from other people. Research on distance education distinguishes among interactions between students and content, students and instructors, and students and other

students (Bernard et al., 2009). Although standalone programs place most of the instructional responsibility on the program, students may still interact with teachers or peers while using them. These interactions can provide additional guidance, encouragement, or clarification, regulation that students may not supply themselves. However, interaction could pull attention away from the program's core instruction, especially if it is poorly built into the program. Our interaction moderator examines whether program effects differ when students interact with instructors or peers. We also examine whether an instructor facilitates program use, such as by monitoring progress or helping students when they encounter difficulties.

The amount of time students spend with a program may also matter. Theories of instructional time suggest that learning depends in part on the time students spend engaged with instruction (Carroll, 1963), so programs used more intensively or over a longer period may provide greater opportunities for learning. We examine intervention duration, weekly hours of use, and frequency of use as moderators.

Finally, the effectiveness of standalone programs may depend on who uses them and where they are used. Younger students may have less developed self-regulation skills, which could make learning independently from a program more difficult. Although they may also benefit from the structure and immediate feedback that some programs provide. We therefore examine grade level without predicting the direction of the association. The value of a program also depends on what students would otherwise receive. Where classrooms are crowded and teachers are stretched, a well-designed program may add more than it does where regular instruction is already strong (Escueta et al., 2020; Muralidharan et al., 2019). We examine country income level and urbanicity as indicators of these broader contexts. We also examine where students access the program. Most programs in our sample were used at school during the

school day, but programs used outside school may place different demands on students and the resources available to support their learning.

Prior Reviews and the Need for a Meta-Analysis

The research on digital learning is extensive, but existing reviews do not fully answer which standalone programs improve PreK–12 mathematics and science achievement or why their effects vary. Many reviews focus on one type of technology, such as computer-assisted instruction, intelligent tutoring systems, or simulations. Others focus on a narrow grade range, or on blended models in which technology supplements a teacher-led instruction (Cheung & Slavin, 2013; Kulik & Fletcher, 2016; Means et al., 2013). Reviews also differ in the studies they include and the approaches they use to synthesize findings, making their estimates difficult to compare. Few reviews separate programs that replace teacher instruction from programs that add to it or compare standalone programs directly with regular classroom instruction. Most importantly for the present study, prior reviews do not consistently distinguish programs in which the technology delivers instructional content from those in which technology supplements instruction delivered by a teacher. Many also emphasize average effects without examining whether differences in program features, implementation, students, and settings help explain variation in outcomes. A synthesis focused specifically on standalone programs can help address these gaps.

This study makes two contributions. First, we systematically review rigorous evaluations of standalone digital learning programs designed to improve PreK–12 mathematics and science achievement. Second, we meta-analyze these evaluations to estimate the average effect of standalone programs compared with traditional classroom instruction and examine variation in effects using moderator analyses and meta-regression. We consider program features,

implementation conditions, student and setting characteristics, and study methods to better understand when and under what conditions these programs are most effective.

For this review, we define a standalone digital learning program as a student-facing program that (a) delivers mathematics or science instructional content, (b) includes at least one core instructional function, such as explanation, guided practice, feedback, adaptive sequencing, or simulation-based exploration, (c) delivers the instructional content without requiring a teacher to present that content, and (d) is evaluated against traditional in-person classroom instruction that does not include the program. We exclude fully online and virtual charter schools, where the intervention is a broader schooling model rather than a particular instructional program (Ahn & McEachin, 2017), as well as blended models in which the teacher remains the main source of instructional content.

Our focus on standalone programs also has a practical origin. We began this work during the COVID-19 pandemic, when widespread school closures created an urgent need to understand how well students could learn through digital programs outside the classroom. Our original aim was therefore to synthesize evaluations of programs delivered remotely. We soon found that few rigorous evaluations tested programs used entirely outside school. Most evaluated programs were used in schools, often in classrooms or computer labs, but still delivered instructional content directly to students. We therefore defined eligible programs according to who delivered the instructional content rather than where students used the program. As a result, this review is not primarily a study of remote learning. It examines standalone digital programs that deliver mathematics or science instruction directly to students, most often as part of their learning time in school.

Research Questions

We address two research questions: (1) What is the average effect of standalone digital learning programs on PreK–12 students' mathematics and science achievement, compared with traditional classroom instruction? (2) How do effects vary with study and method characteristics, program characteristics (e.g., dosage, personalization, instructional practices, and interaction), sample characteristics, setting characteristics, and outcome characteristics?

Methods

Literature Search

We worked with a research librarian to develop a search strategy covering four concepts: the PreK–12 population, digital and remote delivery, mathematics and science content, and evaluation designs. We combined search terms for these concepts using Boolean operators and searched titles and abstracts. We searched EBSCO (including ERIC, APA PsycINFO, and Academic Search Premier), Web of Science Core Collection, and ProQuest Dissertations & Theses Global, limiting results to English-language studies published from January 2007 through December 2022. Full search strategies are provided in Supplemental Material S1.

To identify unpublished and difficult-to-locate studies, we also manually searched the websites of 23 research organizations and review libraries (Supplemental Material S2). In addition, we examined the reference lists of eligible studies and relevant reviews and searched for studies that cited eligible studies. The database searches identified 6,061 records, of which 5,887 remained after duplicates were removed. These additional search methods identified another 149 records, resulting in 6,036 unique records for screening.

We excluded 4,927 records during title and abstract screening and assessed 1,092 full-text reports for eligibility. Of these, 606 met the broader inclusion criteria, and 87 studies met the eligibility criteria for the present review and proceeded to coding. During coding, we excluded an additional 19 studies, most commonly because they did not use random assignment, had a

confounded design, reported only subsample results, or did not provide sufficient information to calculate an effect size. The final analytic sample included 68 studies reported across 81 documents and contributing 147 effect sizes (Figure 1).

Screening

We screened titles and abstracts in Abstrackr, a free, web-based screening tool (Wallace et al., 2012). Each record was judged against five questions. Was the sample in grades PreK–12? Did the study report quantitative results from primary research? Did it include a comparison group? Did the program deliver mathematics or science content? Was the program a student-focused web- or app-based tool? Records were excluded if any answer was clearly "no" and records with unclear answers moved forward to full-text review.

Before screening began, all team members screened the same 100 abstracts to calibrate application of the criteria and train new screeners. Two researchers then screened each abstract independently. Abstrackr uses active learning to prioritize records predicted to be most relevant based on prior screening decisions. Each week, two screeners resolved disagreements, with difficult cases discussed by the full team. After approximately 10 weeks of double screening, Abstrackr predicted that no relevant records remained among the unscreened records. At that point, 2,272 records (39%) had been screened independently by two reviewers, and 38% had advanced to full-text review. We then switched to single screening for the remaining 3,615 records, which were divided among five team members. Of these records, 2.5% advanced to full-text review.

At the full-text stage, we assessed studies against the eligibility criteria specified in the broader review protocol. These criteria covered grade level, use of a student-focused, web- or app-based program delivering mathematics or science content, primary quantitative research, at least one mathematics or science achievement outcome, an eligible comparison group, and an

eligible study design. Each full-text report was initially screened by one trained reviewer. A second reviewer validated a subset of decisions, with an emphasis on exclusions because incorrectly excluded studies would receive no further review. Overall, a second reviewer validated 365 of the 1,065 screening decisions (34%), including 329 of the 463 exclusions (71%). Reviewers discussed disagreements and revised decisions when appropriate, and unclear cases were brought to weekly team meetings.

Because the broader review was originally designed to examine remote education programs, its eligibility criteria were broader than those used for the present meta-analysis. The protocol permitted quasi-experimental designs, a wider range of comparison conditions, and programs that could be used remotely. We therefore conducted an additional eligibility review to identify the standalone digital learning programs examined in this paper. We first limited the sample to randomized controlled trials in which the comparison group received traditional in-person classroom instruction without the program. We then examined how instruction was delivered within each program. Programs were eligible when the digital program delivered the mathematics or science content directly to students without requiring a teacher to present that content. A teacher could be present while students used the program, but the teacher's intended role during program use was limited to supporting students rather than delivering the instructional content. Programs that supplemented teacher-led instruction and blended models in which the teacher remained the primary source of content were excluded.

We also recorded where students were relative to the instructor while using the program: outside school or in a different building, in the same building but a different room, in the same room with an instructor who did not deliver content, or with no instructor present. These distinctions allowed us to separate where students used a program from who delivered the

instruction, with the latter determining eligibility for this review. Only studies meeting the definition of standalone digital instruction proceeded to full coding. Figure 1 presents the number of studies excluded at this stage and the reasons for exclusion.

Coding

We coded studies in MetaReviewer, a free, web-based platform for systematic reviews and meta-analyses (Polanin et al, 2026). The coding protocol covered study and method characteristics, intervention and comparison conditions, sample and setting characteristics, outcome characteristics, and information needed to calculate effect sizes. The full coding protocol is available on the Open Science Framework (https://osf.io/hf94y/overview?view_only=1aadfdd0161b4be1ac8fbda40502c3d6). To calibrate coding, the full team first coded two studies together and discussed disagreements. Researchers then coded the remaining studies independently, with one primary coder assigned to each study. An experienced researcher validated every coded study against the full text and either approved the coding or returned it to the coder for revision. When studies did not report sufficient information to calculate an effect size, we contacted the authors.

We coded moderators corresponding to the program features and conditions described in our conceptual framework (see Table 2 for categories and descriptive statistics). Study and method characteristics included peer-review status, reported conflicts of interest, level of random assignment, use of researcher-developed outcome measures, and whether the program targeted mathematics or science.

Program characteristics included duration in weeks, hours of use per week, and frequency of use. To capture personalization, we coded whether programs provided feedback on students' responses and whether they adapted content to students' skill levels. We also coded four instructional practices: practice or rehearsal (PR), direct instruction (DI), collaborative learning

(CL), and learning through discovery (LD). We combined these practices into five instructional profiles: PR only, CL or LD only, PR with CL or LD, PR with CL or LD and DI, and PR with DI. We also coded whether an instructor facilitated program use, whether students interacted with teachers or peers through the program, and where students accessed the program: at home, at school during the school day, or at school outside school hours.

Sample and setting characteristics included grade level and the proportions of students who were female, from each racial and ethnic group, receiving special education services, English learners, and economically disadvantaged. Setting characteristics included country income level, based on the World Bank classification for fiscal year 2025, and urbanicity.

Outcome characteristics included the mathematics or science subdomain, how the outcome was administered or reported (standardized or state test, researcher-administered measure, or teacher-administered measure or course grade), and the timing of measurement. We classified outcomes as immediate when they were measured during the intervention or directly afterward and as follow-up when they were measured later. Grade level, urbanicity, personalization features, and interaction types were coded as separate indicators when categories could overlap. For example, a study could include students from multiple grade bands, and a program could provide both feedback and adaptive content.

For each eligible achievement outcome, we coded the information needed to calculate an effect size, along with information about the study design and analysis needed for effect size estimation. We coded every eligible achievement outcome reported in each study. When a study reported results for both the full sample and subgroups, we used the full-sample results.

Effect Size Computation

We computed effect sizes following What Works Clearinghouse (WWC) procedures (WWC, 2022). We expressed effects as Hedges' *g*, a standardized mean difference corrected for

small-sample bias, with positive values indicating better performance for students in the program group than for those in the comparison group. For outcomes on which lower scores indicated better performance, we reversed the sign of the effect size.

Because studies reported results in different forms, we used the most informative statistics available to calculate each effect size. When studies reported covariate-adjusted posttest means or a regression coefficient for the treatment effect, we standardized the adjusted difference using the pooled posttest standard deviation. When only unadjusted means were available, we calculated the standardized posttest difference. If pretest means were also available, we adjusted for baseline differences by subtracting the pretest difference from the posttest difference. When studies reported only test statistics, we converted *t* statistics and two-group *F* statistics to standardized mean differences. Because *F* statistics do not indicate the direction of an effect, we determined the direction from the reported results. For dichotomous outcomes, we used the Cox index to convert the log odds ratio to the standardized mean difference scale. For three studies, the reported information was insufficient to independently calculate an effect size, so we used the standardized effect size reported by the study authors.

Some studies randomly assigned classrooms or schools but analyzed student-level outcomes without accounting for the dependence of students within clusters. For these studies, we applied the WWC adjustment for clustering using an intraclass correlation coefficient of .20 for achievement outcomes (WWC, 2022). We incorporated this adjustment into the effect size calculations, sampling variances, and degrees of freedom used for the small-sample correction.

Handling Missing Data

Study, method, and outcome characteristics were complete, whereas missingness in program, sample, and setting characteristics ranged from 4% to 47% across moderators, most often because studies did not report the relevant information (Table 2). To retain effect sizes with

missing moderator data, we used multiple imputation (Pigott, 2019). We imputed program, sample, and setting characteristics in separate blocks using multilevel models that accounted for effect sizes nested within studies. Each imputation model included the variables in that block, study and method characteristics, and the effect size. We conducted the imputations using the R package *jomo* (Quartagno et al., 2019), creating 20 imputed datasets for each block. We fit each analysis across the 20 datasets and pooled estimates using Rubin's rules (Rubin, 1987). Several student demographic characteristics were reported primarily by U.S. studies and too infrequently to support multiple imputation. We therefore examined these characteristics separately in exploratory analyses restricted to U.S. studies that reported the relevant characteristic (Supplemental Material S6).

Meta-Analysis and Model Building

Most studies contributed multiple effect sizes, such as results for several outcomes or follow-up points from the same sample. To account for dependence among these estimates, we used a correlated and hierarchical effects model (Pustejovsky & Tipton, 2022). The model included random effects representing variation between studies and variation among effect sizes within studies. It also allowed effect sizes from the same sample to have correlated sampling errors. Because studies rarely reported correlations among outcomes, we assumed a sampling-error correlation of .50 and examined alternative values in sensitivity analyses.

We estimated all models using restricted maximum likelihood in the R package *metafor* (Viechtbauer, 2010). We calculated cluster-robust standard errors using the CR2 small-sample correction and Satterthwaite degrees of freedom, as implemented in the R package *clubSandwich* (Pustejovsky & Tipton, 2018; Tipton, 2015). Because robust tests may be unreliable when the Satterthwaite degrees of freedom are below 4, we flag these results and interpret them cautiously (Tipton, 2015).

We first fit a model without moderators to estimate the average effect and the variation in effects between and within studies. We then examined moderators in blocks corresponding to our research questions. The first block included study and method characteristics, which we treated as controls because they describe how studies were conducted rather than the programs themselves. We examined each of the other four blocks, program, sample, setting, and outcome characteristics, in a separate model that also included the study and method characteristics. We log-transformed duration and intensity because their distributions were highly skewed. We entered categorical moderators as indicators relative to a reference category. For moderators whose categories could overlap, we entered each category as a separate indicator.

For moderators represented by multiple indicators, we used Wald tests to evaluate whether the indicators were jointly associated with effect sizes. For analyses involving imputed moderator data, we pooled the Wald statistics across imputations. Because no established method is available for pooling robust Wald tests with small-sample corrections across multiply imputed datasets, we used $m - 1$ denominator degrees of freedom, where m was the number of studies (Pustejovsky & Tipton, 2018). For tests of individual coefficients involving imputed data, we used the smaller of the Rubin and Satterthwaite degrees of freedom.

We next fit a joint model that included the study and method characteristics together with selected moderators from the other blocks: program intensity, personalization features, instructional practices, country income level, and science subdomain. We selected these moderators based on the results of the block models. Because this selection was data-driven, the joint model was not intended to identify a definitive set of predictors or provide an independent confirmatory test. Instead, we used it to examine whether patterns observed in the block models remained when the selected moderators were considered together and to estimate the amount of

heterogeneity they explained jointly. We summarized the remaining heterogeneity as the square root of the sum of the estimated between-study and within-study variance components.

Finally, we used MetaForest as an exploratory check on the meta-regression findings. MetaForest applies random forests to meta-analytic data, allowing for nonlinear relationships and interactions without requiring these relationships to be specified in advance (Van Lissa, 2020). We used it to rank moderators according to their contribution to the prediction of effect sizes, not to test their statistical significance. Supplemental Material S7 provides details about this analysis.

Interpreting Results

To make the moderator results easier to interpret, we report conditional means rather than regression coefficients for categorical moderators. These conditional means represent the predicted average effect size for each category, with the other moderators in the model held at their sample means. This puts all categories on the same effect-size scale and allows readers to compare predicted effects directly rather than interpreting differences relative to a reference category. For continuous moderators, such as program intensity, we report the regression slope. For analyses involving imputed data, we calculated conditional means within each imputed dataset and pooled the estimates using Rubin's rules.

Some moderators, including grade level, urbanicity, personalization features, and types of interaction, were represented by indicators that were not mutually exclusive. For these moderators, the regression reference group consisted of studies for which none of the indicators was present. For example, the reference group for personalization was programs that provided neither feedback nor adaptive content. Conditional means for each indicator represent the predicted average effect when that feature is present, with the other indicators and moderators held at their sample means.

For categorical moderators, the reported *p* value is based on the overall test of whether effects differ across categories; it does not test whether an individual conditional mean differs from zero. As noted earlier, we flag tests with fewer than 4 Satterthwaite degrees of freedom because robust inference is unreliable with so few degrees of freedom. We therefore treat these estimates as descriptive rather than inferential.

Sensitivity Analyses and Selective Reporting Bias

We checked how much the average effect depended on our modeling choices and on individual studies. First, we refit the model with no moderators using assumed correlations of 0, 0.2, and 0.8 among effect sizes from the same sample, in place of 0.5. Second, we estimated a confidence interval for the average effect using a cluster bootstrap. We resampled studies 1,000 times and computed bias-corrected and accelerated intervals, which do not rely on the small-sample approximations used in robust variance estimation. Third, we refit the model 68 times, leaving out one study each time, to see whether any single study drove the result. We also computed Cook's distance for each study and flagged studies with values above $4/m$, where m is the number of studies (Viechtbauer & Cheung, 2010).

Selective reporting can bias a meta-analysis if studies with statistically significant results are more likely to be published or reported. Our search included dissertations and other unpublished reports, and peer review status was one of the method characteristics we tested as a moderator. We also used two formal methods. The first was a three-parameter selection model (Vevea & Hedges, 1995), which estimates the average effect while allowing studies with nonsignificant results to have a different chance of being reported than studies with significant positive results. We used one cut point at a one-tailed *p* value of 0.025. We estimated the model with the R package *metaselection*, which accounts for multiple effect sizes per study, and computed confidence intervals using 1,000 cluster bootstrap resamples (Pustejovsky et al.,

2026). The second was a funnel plot with the trim-and-fill method (Duval & Tweedie, 2000). Trim and fill assumes that effect sizes are independent, so we treat it as descriptive only. Results of all sensitivity and selective reporting analyses are reported in Appendix Table A2.

Results

Characteristics of Included Studies

Table 2 summarizes the characteristics of the included studies, programs, samples, and outcomes. The 68 studies were conducted across 23 countries, including 24 in the United States. Most were conducted in high-income countries (63%), published in peer-reviewed journals (81%), and evaluated mathematics programs (72%). Most studies randomly assigned individual students (62%), and 15% reported a conflict of interest, most often involving the program developer.

Half of the studies included students in elementary grades, while 25% included middle school students and 35% included high school students; some studies spanned multiple grade bands. Other student characteristics, including race and ethnicity, English learner status, and economic disadvantage, were reported primarily in U.S. studies and were missing for more than 80% of effect sizes.

Programs lasted an average of about 11 weeks (SD = 12) and were used for approximately 3 hours per week (SD = 4). Most provided feedback on students' responses (72%), while 35% adapted content to students' skill levels. The most common instructional approaches combined practice with direct instruction (31%) or used practice alone (29%). Most programs were used at school during the school day (88%), with relatively few studies evaluating programs used at home or outside regular school hours.

Most outcomes were researcher-administered measures (59% of effect sizes), and 88% measured achievement during the intervention or immediately afterward. Only nine studies reported outcomes at a later follow-up.

Meta-Analysis Results

Unconditional model. Across all studies, standalone digital learning programs had a positive average effect on mathematics and science achievement, $g = 0.47$ ($SE = 0.06$, 95% CI [0.35, 0.60], $p < .001$), compared with traditional classroom instruction. The cluster bootstrap produced a similar 95% confidence interval [0.34, 0.63].

Effects varied considerably both between and within studies. The between-study standard deviation was $\tau_B = 0.43$ and the within-study standard deviation was $\tau_W = 0.29$, corresponding to a combined heterogeneity standard deviation of $\tau = 0.52$. The 95% prediction interval ranged from $g = -0.54$ to 1.49, indicating substantial variation in the effects that might be observed across studies and settings. Based on the estimated distribution of true effects, 82% were estimated to exceed zero (95% CI [75%, 88%]), 76% to exceed $g = 0.10$, and 67% to exceed $g = 0.25$. The heterogeneity estimates from this model provide the baseline for assessing how much variation is accounted for by moderators.

Study and Method Characteristics. We first fit a base model containing study and method characteristics (Table 3), which were retained as controls in the subsequent moderator models. Conflict of interest showed the largest difference. Studies with an identified conflict of interest, most often involving the program developer, had larger conditional mean effects than studies without an identified conflict ($g = 1.01$ vs. 0.39), although the difference did not reach the conventional threshold for statistical significance ($p = .06$). The other study and method characteristics were not associated with effect sizes. Effects were similar for peer-reviewed and non-peer-reviewed studies ($g = 0.46$ vs. 0.49, $p = .90$), across levels of random assignment ($p =$

.77), and for studies with and without researcher-developed measures (g = 0.47 vs. 0.46, p = .93). Effects were also similar for mathematics and science programs (g = 0.46 vs. 0.49, p = .84).

Program Characteristics. Several program characteristics were associated with differences in effect sizes in the block model (Table 4). Programs used for more hours per week had larger effects (b = 0.19 per unit increase in log hours, p = .045). This corresponds to an increase of approximately 0.13 in effect size for a doubling of weekly use. Duration was not associated with effects (b = 0.05, p = .59), nor was frequency of use overall (p = .47).

Personalization showed a suggestive pattern, although the overall test was not significant (p = .11). Programs that provided feedback had a larger conditional mean effect than programs with no personalization (g = 0.53 vs. 0.19), and the feedback contrast was statistically significant (b = 0.36, p = .0497). Adaptivity was not independently associated with larger effects (b = −0.06, p = .77). Because most adaptive programs also provided feedback, the conditional mean for adaptivity (g = 0.41) largely reflects programs that included both features.

Instructional practices showed a similar pattern. Programs combining practice with direct instruction had the largest conditional mean effect (g = 0.74) and a larger effect than programs using practice alone (g = 0.22; b = 0.52, p = .02). However, the overall test across the five instructional profiles was not significant (p = .19). Programs combining practice with collaboration or discovery had conditional mean effects ranging from g = 0.36 to 0.52, while programs based on collaboration or discovery without practice had a conditional mean of g = 0.46. These profiles did not differ significantly from practice alone.

We found less evidence of differences associated with other program characteristics. Conditional mean effects were smaller for programs that allowed interaction with teachers (g = 0.12) or peers (g = 0.27) than for programs without these forms of interaction (g = 0.57), but

neither contrast was statistically significant (both $p = .23$; overall $p = .14$). Effects also did not differ significantly according to whether an instructor facilitated program use ($g = 0.37$ vs. 0.49, $p = .48$) or where students accessed the program ($p = .60$). The latter estimates were based on relatively few studies outside the typical school-day setting: six studies evaluated programs used at home and three evaluated programs used at school outside regular school hours.

Sample Characteristics. Grade level and gender composition were not significantly associated with effect sizes (Table 5). Conditional mean effects were largest for samples limited to elementary grades ($g = 0.56$) and smaller for samples that included middle school ($g = 0.33$) or high school students ($g = 0.34$). Neither contrast was statistically significant (both $p = .19$), and the overall grade-level test was also not significant ($p = .30$). The proportion of female students was not associated with effect sizes ($b = 0.41$ per unit change in proportion, $p = .42$).

Race and ethnicity, English learner status, economic disadvantage, and other student characteristics were reported primarily by U.S. studies. Exploratory analyses restricted to U.S. studies found no reliable associations with effect sizes (Appendix Table A1). These analyses included only 7 to 11 studies each, and most had fewer than 4 degrees of freedom. We therefore treat these results as descriptive.

Setting Characteristics. Country income level was associated with effect sizes (Table 6; overall $p = .002$). Programs evaluated in high-income countries had a smaller conditional mean effect ($g = 0.28$) than programs evaluated in low- or lower-middle-income countries ($g = 0.76$) or upper-middle-income countries ($g = 0.84$). The contrast between high-income and low- or lower-middle-income countries was statistically significant ($b = -0.48$, $p = .04$), while effects in upper-middle-income countries did not differ significantly from those in low- or lower-middle-income countries ($b = 0.08$, $p = .76$). Urbanicity was not associated with effect sizes ($p = .99$).

Conditional mean effects were similar across urban (g = 0.46), rural (g = 0.43), and suburban (g = 0.46) settings.

Outcome Characteristics. Mathematics effects did not differ across subdomains (Table 7; p = .94). Conditional mean effects were similar for general or mixed mathematics (g = 0.50), number and operations (g = 0.45), and algebra, geometry, and data (g = 0.46).

Effects varied more across science subdomains, although the overall test was not statistically significant (p = .07). Conditional mean effects were close to zero for biology and life science outcomes (g = 0.05) and larger for physical sciences (g = 0.60) and general, earth, or other science outcomes (g = 0.65). Relative to biology and life science, the contrast was statistically significant for general, earth, or other science (b = 0.60, p = .04) and was not statistically significant for physical sciences (b = 0.55, p = .09). These estimates should be interpreted cautiously because each science subdomain was represented by only 6 to 8 studies.

Effects did not differ according to how outcomes were measured (p = .92). Conditional mean effects were similar for researcher-administered measures (g = 0.45), standardized or state tests (g = 0.51), and teacher-administered measures or course grades (g = 0.44). Effects measured at later follow-up (g = 0.43) also did not differ significantly from those measured during or immediately after the intervention (g = 0.47, p = .81). Only nine studies reported later follow-up outcomes, however, limiting our ability to draw conclusions about the persistence of effects.

Final Model and Variance Explained. The final model combined the study and method characteristics with moderators selected from the block models: intensity, personalization features, instructional practices, science subdomain, and country income level (Table 8). As

described in the Methods, this model was used to examine whether patterns from the block models remained when these moderators were considered together.

Country income level remained strongly associated with effects (overall $p < .001$). Programs evaluated in high-income countries had a conditional mean effect of $g = 0.24$, compared with $g = 0.87$ in low- or lower-middle-income countries and $g = 0.86$ in upper-middle-income countries. The contrast between high-income and low- or lower-middle-income countries remained statistically significant ($b = −0.62$, $p = .01$).

Science subdomain was also associated with effects in the final model (overall $p < .01$). Conditional mean effects remained near zero for biology and life science outcomes ($g = −0.02$) and were larger for physical sciences ($g = 0.69$; $b = 0.71$, $p = .01$) and general, earth, or other science outcomes ($g = 0.61$; $b = 0.63$, $p = .007$). Because each science subdomain was represented by relatively few studies, these differences should be interpreted cautiously.

Patterns for program characteristics were generally similar to those in the block model, but the statistical evidence was weaker. Personalization was marginally associated with effect sizes overall ($p = .08$). Programs with no personalization had the smallest conditional mean effect ($g = 0.14$), compared with programs providing feedback ($g = 0.51$) or adaptive content ($g = 0.62$), although neither individual contrast was statistically significant ($p = .12$ and .09, respectively). Practice combined with direct instruction again had the largest conditional mean among the instructional profiles ($g = 0.62$), but neither its contrast with practice alone ($g = 0.40$, $p = .11$) nor the overall test ($p = .27$) was statistically significant. Intensity was also no longer associated with effects after the other moderators were included ($b = 0.05$, $p = .40$).

Among the study and method characteristics, the conflict-of-interest difference remained similar to that observed in the base model ($g = 0.88$ vs. 0.39, $p = .06$). Level of random

assignment was associated with effects in the final model (p = .048), with larger conditional mean effects in studies assigning teachers or classrooms (g = 0.69) than in studies assigning students (g = 0.41) or schools (g = 0.32). Because the final model also includes science subdomain indicators, its conditional mean for science programs largely reflects the omitted science subdomain. Table 3 therefore provides the more appropriate overall comparison between mathematics and science programs.

Table 9 summarizes heterogeneity across models. The base model reduced the combined heterogeneity standard deviation by 2.1% relative to the unconditional model. Relative to the base model, the largest reductions were observed after adding setting characteristics (10.7%) and program characteristics (8.2%), while the reductions associated with sample (1.1%) and outcome characteristics (0.2%) were small. In the final model, the combined heterogeneity standard deviation decreased from $\tau = 0.52$ in the unconditional model to $\tau = 0.39$, a reduction of 24.8%. This reduction occurred primarily in between-study heterogeneity, which declined from τ_B = 0.43 to 0.26. Within-study heterogeneity remained approximately $\tau_W = 0.29$ across models. Thus, the moderators included in the final model accounted for some of the differences in effects across studies but little of the variation among outcomes within studies.

The exploratory MetaForest analysis produced a broadly similar pattern (Supplemental Material S7). Country income level had the highest variable-importance ranking, followed by conflict of interest, instructional practices, intensity, and level of random assignment. Moderators showing little evidence of association in the meta-regression generally received lower importance rankings.

Sensitivity Testing and Selective Reporting Bias

The average effect was robust across the sensitivity analyses (Appendix Table A2). Varying the assumed sampling-error correlation among effect sizes from the same sample had

little effect on the results. Across assumed correlations ranging from 0 to .80, the average effect ranged from $g = 0.47$ to 0.48, and the combined heterogeneity standard deviation ranged from $\tau = 0.51$ to 0.52. The cluster bootstrap produced a 95% confidence interval of [0.34, 0.63], similar to the interval from the primary model. Results were also stable when individual studies were omitted. Across the 68 leave-one-study-out models, the average effect ranged from $g = 0.44$ to 0.49 and remained statistically significant in every case. Four studies had Cook's distance values greater than $4/k$, where k is the number of studies. All four reported relatively large positive effects, but omitting any one of them reduced the average effect by only 0.02 to 0.04.

The contour-enhanced funnel plot showed some asymmetry (Figure S2). Among less precise estimates, effect sizes were mostly positive, and a few small studies reported very large effects. This pattern could reflect selective reporting, but it could also arise if smaller studies differed systematically from larger studies. The formal analyses did not provide clear evidence that selective reporting explained the pattern. The three-parameter selection model estimated an average effect of $g = 0.49$ (95% CI [0.31, 0.70]), similar to the primary estimate. The estimated relative probability of observing nonsignificant results compared with significant positive results was 1.5 (95% CI [0.70, 3.17]). Because the confidence interval included 1, the analysis did not indicate that nonsignificant results were less likely to be observed. The trim-and-fill method did not impute any missing effect sizes, although this analysis is descriptive because it assumes independent effect sizes. Peer-reviewed and unreviewed studies also had similar effects (Table 3).

Discussion

This review asked whether students can learn mathematics and science from digital programs that deliver instruction directly, without requiring a teacher to present the content.

Across 68 randomized controlled trials, the answer was generally yes. Students who used standalone digital learning programs outperformed students receiving traditional classroom instruction by an average of $g = 0.47$. Effects varied considerably across studies, indicating that standalone programs are not uniformly effective. Even so, the estimated distribution of true effects was largely positive: about four in five effects were estimated to exceed zero, and about two in three to exceed $g = 0.25$.

The average effect in this review is larger than estimates reported in some earlier reviews of educational technology. Cheung and Slavin (2013), for example, reported an average effect of approximately 0.16 for technology applications in K–12 mathematics, many of which supplemented teacher-led instruction. These estimates are not directly comparable, however, because the reviews differ in the studies they included, the settings represented, and the methods used to estimate effects. Our review included only randomized trials, relied more heavily on researcher-administered measures, and included a substantial number of studies conducted outside the United States. Still, the findings challenge the assumption that digital programs are useful primarily when they supplement teacher-led instruction. Standalone programs can improve mathematics and science achievement when they deliver instructional content directly to students. The more useful question, then, is not simply whether these programs can work, but which programs work best, for which students and settings, and under what conditions. The moderator analyses provide some initial answers to these questions.

Developer Involvement

The conflict-of-interest finding deserves attention because developer involvement has long been a concern in educational technology research. Studies with an identified conflict of interest, most often evaluations conducted or funded by the program developer, reported effects more than twice as large as studies without an identified conflict ($g = 1.01$ vs. 0.39). Although

the difference did not reach the conventional threshold for statistical significance (p = .06) and was based on only 10 studies with an identified conflict, the pattern was similar in the block and final models. It is also consistent with prior research showing larger effects in developer-involved or developer-commissioned evaluations than in independent evaluations (Cheung & Slavin, 2016; Wolf et al., 2020).

Several factors could contribute to this difference. Developer-involved studies may use outcome measures that are more closely aligned with the program, implementation may receive greater support or oversight, or favorable findings may be more likely to be reported. Our data cannot distinguish among these explanations. Importantly, however, studies without an identified conflict of interest still showed a positive average effect (g = 0.39). Thus, the overall finding that standalone digital programs can improve student achievement is not limited to studies with identified developer involvement. At the same time, the size of the difference highlights the value of independent evaluations when schools and other decision makers assess evidence for a particular program.

Which Program Features Mattered

Several patterns in the moderator analyses were consistent with the learning theories that motivated them, although the evidence for individual program features was stronger in the block models than in the final model. One of the clearest involved feedback. Consistent with research on tutoring and feedback (Hattie & Timperley, 2007; VanLehn, 2011), programs that responded to students' answers had larger effects than programs that offered no personalization (g = 0.53 vs. 0.19 in the block model). Adaptivity was not independently associated with larger effects once feedback was taken into account. Because most adaptive programs also provided feedback, however, the two features were difficult to disentangle. We also coded only whether a program provided feedback, not the type or quality of that feedback. Future research could distinguish

explanatory feedback from simpler information about whether a response was correct, a distinction that prior research suggests may matter for learning (Wisniewski et al., 2020).

Instructional practices showed another potentially important pattern. Programs that combined practice with direct instruction had larger effects than programs that relied on practice alone (g = 0.74 vs. 0.22 in the block model). This pattern is consistent with the idea that when a program delivers new material directly to students, explanations and worked guidance may be particularly important (Kirschner et al., 2006). Programs that added collaboration or discovery to practice did not show reliably larger effects than practice alone, and programs centered on collaboration or discovery fell between these groups. These findings are broadly consistent with research suggesting that discovery learning benefits from guidance (Alfieri et al., 2011). At the same time, the overall test across instructional profiles was not significant, and the contrast between practice with direct instruction and practice alone weakened in the final model. These patterns should therefore be viewed as suggestive rather than definitive evidence that particular instructional approaches are more effective.

The findings for student support and self-regulation were less clear. Programs that allowed students to interact with teachers or peers had smaller estimated effects than programs without these features, but the differences were not statistically significant. Instructor facilitation was also unrelated to effects. Grade level showed no significant association, and there was no indication that standalone programs were less effective for younger students. In fact, effects were largest for samples limited to elementary grades, although the differences across grade levels were not statistically significant. This pattern is consistent with evidence that well-designed digital programs can support younger students' mathematics learning (Kim et al., 2021), but our

analyses cannot show whether features such as structure, pacing, or feedback compensated for differences in students' self-regulation.

Finally, programs used for more hours per week had larger effects in the block model. This pattern is consistent with theories emphasizing time engaged in instruction (Carroll, 1963) and with dose-response findings from experimental studies of computer-assisted learning (Muralidharan et al., 2019). The association weakened substantially in the final model, however, suggesting that intensity may overlap with other characteristics of programs or the settings in which they are implemented. These results point most consistently toward feedback, instructional support, and intensity as potentially important features, but they do not establish that any of these features independently causes larger effects.

Where Programs Were Used

The strongest moderator in our analyses was the income level of the country where a program was evaluated. Programs evaluated in low- and middle-income countries produced average effects of roughly 0.8, compared with about 0.25 in high-income countries, and this pattern remained strong across models. One possible explanation is that the value of a digital program depends partly on the instruction and resources students would otherwise receive. In settings where instructional resources are more limited, a program that provides structured instruction, practice, and immediate feedback may offer more than it does in settings where students already have access to strong instructional supports. Prior research on educational technology in low- and middle-income countries provides some support for this explanation, particularly for programs that allow students to work at an appropriate level and provide individualized learning opportunities (Ganimian & Murnane, 2016; Muralidharan et al., 2019; Rodriguez-Segura, 2022).

Country income level, however, captures many differences beyond the strength of regular classroom instruction. Studies conducted in different countries may evaluate different kinds of programs, serve different student populations, measure different outcomes, or differ in how programs are implemented. Income level was also related to some program features in our data, including adaptivity, and our moderator analyses could adjust only for characteristics that we were able to code. The income finding is therefore best understood as a robust pattern whose explanation remains uncertain, rather than evidence that the same program would necessarily produce a larger effect in a lower-income setting. Urbanicity, a narrower measure of context, was not associated with effect sizes. These findings suggest that the benefits of standalone digital learning programs may depend substantially on the broader educational context in which they are used. The smaller average effect in high-income countries, about 0.25, also provides a more relevant benchmark for decision makers considering these programs in those settings.

Measurement and Outcomes

Effects were similar across different types of outcome measures. Researcher-administered measures, standardized or state tests, and teacher-administered measures or course grades produced similar estimated effects. This is notable because earlier reviews have found larger effects on researcher-developed measures, which may be more closely aligned with the content of an intervention (Cheung & Slavin, 2016; Kim et al., 2021). We did not observe that pattern in this review, providing some reassurance that the overall effect was not driven primarily by the type of measure used.

Science outcomes showed a different pattern. Effects were close to zero for biology and life science outcomes and larger for physical sciences and general, earth, or other science outcomes. The overall difference across science subdomains was not statistically significant in the block model but became significant when other moderators were included in the final model.

Because each science subdomain was represented by only six to eight studies, these differences should be interpreted cautiously. We cannot determine whether they reflect differences in the content itself, the types of digital programs used to teach it, or other characteristics of the studies. Still, the pattern suggests that the effectiveness of standalone digital instruction may differ across areas of science and warrants further study.

Implications

Our findings should be read in light of how these programs were used. Most were not designed to replace teachers or schooling, and in most studies students used them at school, for a limited part of their instructional time, while continuing to receive regular instruction otherwise. What our review isolates is the effect of letting a program, rather than a teacher, deliver instruction during that time. The results suggest that well-designed programs can do this effectively. They do not show that programs can substitute for teachers more broadly or support fully remote learning, which few studies tested.

Within those bounds, the choice of program matters for school and district leaders. Programs that give students feedback on their answers and that pair practice with direct instruction were associated with the largest effects. Programs that offer practice alone, without explanation, were associated with the smallest. School leaders evaluating a product can ask whether it explains concepts as well as assigning problems, and whether it responds to students' mistakes. They can also plan for enough weekly use, since programs used more intensively tended to produce larger gains. Where possible, they should look for independent evaluations, given the larger effects reported in studies with a conflict of interest.

For program developers, the results point to the value of building instruction into the program rather than relying on a teacher to provide it. Clear explanations, worked examples, and feedback that responds to each answer appear to matter more than adaptive sequencing on its

own. Exploratory and collaborative features may need explicit guidance built into the program to be effective when students work without a teacher.

For researchers, the review highlights several gaps. Few rigorous studies evaluated programs used at home or outside school hours, even though remote use motivated much of the interest in these programs during the pandemic. Many studies did not report basic information about their samples, such as students' race and ethnicity, English learner status, or economic disadvantage, which limited our ability to examine whether programs work equally well for different groups of students. Studies also described program features and implementation unevenly. More consistent reporting of how programs were designed and used, how much students actually used them, and how outcomes were measured would allow future reviews to explain more of the variation in effects. Finally, independent evaluations of widely used products, particularly in high-income countries where average effects were smaller, would help schools make better-informed decisions.

Limitations

Several limitations should be kept in mind. First, as noted above, most programs in our review were used at school for part of students' instructional time, and only a handful of studies tested programs used at home. Our findings speak to what programs can do when they deliver instruction within a school day, not to fully remote learning. Second, our moderator analyses are correlational. Program features, settings, and study methods were not randomly assigned across studies, and some were related to one another. For example, adaptivity was more common in some income groups than others. Although each block model adjusted for study and method characteristics, the moderator results describe which kinds of programs and settings were associated with larger effects, not the causal effect of any single feature.

Third, several analyses rested on small numbers of studies, and some moderator results were significant for individual contrasts but not for the overall test. With many moderators tested, some of these results may be chance findings, which is why we emphasize patterns that held across models and fit prior theory. Fourth, our coding captured whether programs included a feature, not how well it was designed or how much students actually used the program, and studies varied widely in how much of this information they reported. Missing values were imputed under the assumption that they could be predicted from the other coded characteristics. Student demographic characteristics other than gender were reported too rarely to analyze in the full sample.

Finally, aspects of our review process limit what we can claim. Full texts were screened by a single reviewer, with a second reviewer validating a subset of decisions, and studies were coded by a single coder and validated by an experienced researcher. We searched only English-language sources published through 2022, and our search of grey literature was limited to selected research organizations. We also restricted the review to randomized controlled trials. This strengthens the internal validity of the included studies but may exclude evaluations of programs implemented at larger scale, which are more often studied with quasi-experimental designs.

Future Directions and Conclusion

Because our moderator findings are correlational, the features we identified need to be tested directly, and many platforms now make it possible to randomize features within a single program (Heffernan & Heffernan, 2014). These questions are becoming more pressing as the tools change. Our search ended in 2022, before generative artificial intelligence (AI) became widely available in schools, and early experiments with AI tutors are mixed. One trial found meaningful gains from an AI math tutor in a low-income setting (Henkel et al., 2024), while

another found that a general-purpose chatbot harmed later learning unless it was designed to guide students rather than give answers (Bastani et al., 2025). These results echo our own. What made standalone programs effective was not the technology but the instruction built into it.

Standalone digital programs are often viewed with suspicion, as if any use of technology in place of a teacher must come at students' expense. Our review offers a more nuanced picture. When programs delivered mathematics and science instruction for part of students' learning time, students generally learned more than in regular classroom instruction, especially when programs explained concepts and gave feedback, and in settings where regular instruction was most constrained. These programs did not replace teachers or schools, and our findings do not suggest that they could. However, they do suggest that well-designed programs can take on a real share of instruction, and that how a program is designed matters as much as whether one is used.

References

Ahn, J., & McEachin, A. (2017). Student enrollment patterns and achievement in Ohio's online charter schools. *Educational Researcher, 46*(1), 44–57. doi:10.3102/0013189X17692999

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1–18. doi:10.1037/a0021017

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199–209. doi:10.1207/s15326985ep4004_2

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. doi:10.1073/pnas.2422633122

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research, 79*(3), 1243–1289. doi:10.3102/0034654309333844

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733. Doi: 10.1177/016146816306400801

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88–113. doi:10.1016/j.edurev.2013.01.001

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292. doi:10.3102/0013189X16656615

Cheung, A., Slavin, R. E., Kim, E., & Lake, C. (2017). Effective secondary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching, 54*(1), 58–81. doi:10.1002/tea.21338

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. doi:10.1111/j.0006-341X.2000.00455.x

Education Week. (2021). *Map: Coronavirus and school closures in 2019-2020*. Retrieved from https://www.edweek.org

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature, 58*(4), 897–996. doi:10.1257/jel.20191507

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., ... LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics—Physics Education Research, 1*(1), 010103. doi:10.1103/PhysRevSTPER.1.010103

Ganimian, A. J., & Murnane, R. J. (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research, 86*(3), 719–755. doi:10.3102/0034654315627499

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. doi:10.3102/003465430298487

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470–497. doi:10.1007/s40593-014-0024-x

Heinrich, C. J., Darling-Aduana, J., Good, A., & Cheng, H. (2019). A look inside online educational settings in high school: Promise and pitfalls for improving educational opportunities and outcomes. *American Educational Research Journal, 56*(6), 2147–2188. doi:10.3102/0002831219838776

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024). *Effective and scalable math support: Evidence on the impact of an AI-tutor on math achievement in Ghana* (arXiv:2402.09809). Retrieved from https://arxiv.org/abs/2402.09809

Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures matter: A meta-analysis of the effects of educational apps on preschool to grade 3 children's literacy and math skills. *AERA Open, 7*. doi:10.1177/23328584211004183

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. doi:10.1207/s15326985ep4102_1

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

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–19. doi:10.1037/0003-066X.59.1.14

Means, B., Bakia, M., & Murphy, R. (2014). *Learning online: What research tells us about whether, when and how*. New York, NY: Routledge.

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The Effectiveness of Online and Blended Learning: A Meta-Analysis of the Empirical Literature. *Teachers College Record: The Voice of Scholarship in Education*, *115*(3), 1-47. doi: 10.1177/016146811311500307

Polanin, J. R., Austin, M., Peko-Spicer, S., Ezzat, Y., Katragadda, S. A., Michaelson, L. E., Iacocca, M. O., Williams, S., Ebersole, C., Lam, T., & Clements, J. (2026). MetaReviewer (Version 1.6.0) [Computer software]. American Institutes for Research. https://www.metareviewer.org/

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review, 109*(4), 1426–1460. doi:10.1257/aer.20171112

National Center for Education Statistics. (2021). *The Nation's Report Card: 2019 NAEP science assessment*. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2022). *Reading and mathematics scores decline during COVID-19 pandemic: NAEP long-term trend assessment results*. Washington, DC: U.S. Department of Education.

Irwin, V., Wang, K., Tezil, T., Zhang, J., Filbey, A., Jung, J., Bullock Mann, F., Dilig, R., & Parker, S. (2023). *Report on the condition of education 2023* (NCES 2023-144). National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2023144

Pigott, T. D. (2019). Missing data in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 367–382). New York, NY: Russell Sage Foundation.

Pustejovsky, J. E., Citkowicz, M., & Joshi, M. (2026). Estimation and inference for step-function selection models in meta-analyses with dependent effects. *Research Synthesis Methods*, 1–34. doi:10.1017/rsm.2026.10117

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683. doi:10.1080/07350015.2016.1247004

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science, 23*(3), 425–438. doi:10.1007/s11121-021-01246-3

Quartagno, M., Grund, S., & Carpenter, J. (2019). jomo: A flexible package for two-level joint modelling multiple imputation. *The R Journal, 11*(2), 205–228. doi:10.32614/RJ-2019-028

Ritter, S., Kulikowich, J., Lei, P.-W., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In *Proceedings of the 2007 conference on Supporting Learning Flow through Integrative Technologies* (pp. 13–20). IOS Press.

Rodriguez-Segura, D. (2022). EdTech in developing countries: A review of the evidence. *The World Bank Research Observer, 37*(2), 171–203. doi:10.1093/wbro/lkab011

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47*(4), 833–878. doi:10.3102/0002831210367426

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley

Shao, J., Chen, Y., Wei, X., Li, X., & Li, Y. (2023). Effects of regulated learning scaffolding on regulation strategies and academic performance: A meta-analysis. *Frontiers in Psychology, 14*, 1110086. doi:10.3389/fpsyg.2023.1110086

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375–393. doi:10.1037/met0000011

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* (pp. 186–202). London, England: Routledge.

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

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419–435. doi:10.1007/BF02294384

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

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. doi:10.1002/jrsm.11

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 819–824). New York, NY: ACM. doi:10.1145/2110363.2110464

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. doi:10.3389/fpsyg.2019.03087

Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average effect sizes in developer-commissioned and independent evaluations. *Journal of Research on Educational Effectiveness, 13*(2), 428–447. doi:10.1080/19345747.2020.1726537

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. doi:10.1207/s15430421tip4102_2

Tables and Figures

Table 1. Descriptive Characteristics of the Included Studies (68 studies, 147 effect sizes)

Characteristics	*% of studies (number of studies)*	*% of effect sizes (number of effect sizes)*	*% of effect sizes missing the characteristic code*
Study and method characteristics			
Peer review status			
Peer reviewed	81% (55)	84% (124)	0%
Not peer reviewed	19% (13)	16% (23)	0%
Conflict of interest (COI) status			
COI identified	15% (10)	12% (18)	0%
No COI identified	85% (58)	88% (129)	0%
Level of random assignment			
Students	62% (42)	63% (92)	0%
Teachers/classrooms	22% (15)	21% (31)	0%
Schools/centers	16% (11)	16% (24)	0%
Researcher-developed measure			
Used RDM	72% (49)	71% (105)	0%
Did not use RDM	32% (22)	29% (42)	0%
Intervention focus			
Math intervention	72% (49)	77% (113)	0%
Science intervention	28% (19)	23% (34)	0%
Intervention characteristics			
Duration in weeks*	-	10.6 (12.0)	10%
Intensity (hours per week)	-	2.8 (3.8)	23%
Frequency of Interaction			
Low (< weekly)	7% (5)	7% (10)	19%
Moderate (1-2 per week)	31% (21)	31% (45)	19%
High (> 3+ per week)	38% (26)	44% (64)	19%
Personalized instruction supports			
No personalization	19% (18)	16% (31)	4%
Provides feedback	72% (49)	77% (113)	4%
Adapts to learner aptitude/ability	35% (24)	31% (46)	4%
Instructional practices (digital app)			
PR only	29% (20)	36% (53)	0%
CL/LD only	19% (13)	14% (20)	0%
PR + CL/LD	9% (6)	12% (17)	0%
PR + CL/LD + DI	12% (8)	8% (12)	0%
PR + DI	31% (21)	31% (45)	0%
Instructor facilitation			
Not facilitated	74% (50)	67% (99)	0%
Facilitated (fully/partially)	28% (19)	33% (48)	0%
Interactions within digital app			
No interactions	76% (52)	70% (103)	0%
Interactions with teachers	10% (7)	14% (21)	0%
Interactions with peers	21% (14)	25% (37)	0%
Intervention access			
Home	9% (6)	10% (14)	0%
School site, during school day	88% (60)	82% (120)	0%
School site, after school	6% (4)	10% (15)	0%
Sample characteristics			
Proportion Female*	-	0.49 (0.11)	28%
Proportion Black*	-	0.28 (0.22)	85%
Proportion Hispanic*	-	0.39 (0.21)	82%

Proportion Asian*	-	0.16 (0.31)	88%
Proportion White*	-	0.44 (0.30)	88%
Percent SPED*	-	0.20 (0.26)	84%
Percent EL*	-	0.20 (0.21)	89%
Proportion econ. disadvantaged*	-	0.52 (0.25)	84%
Grade level			
Elementary school (PreK-5)	50% (34)	54% (79)	0%
Middle school (6-8)	25% (17)	21% (31)	0%
High school (9-12)	35% (24)	33% (48)	0%
Setting characteristics			
Country income level			
Low/lower-middle	18% (12)	18% (26)	0%
Upper middle	19% (13)	16% (23)	0%
High	63% (43)	67% (98)	0%
Urbanicity			
Rural	21% (14)	22% (33)	47%
Urban	39% (26)	41% (61)	47%
Suburban	12% (8)	7% (11)	47%
Outcome measure characteristics			
Math outcomes			
General or mixed math	29% (20)	29% (43)	n/a
Number and operations	31% (21)	29% (42)	n/a
Algebra, geometry, and data	21% (20)	29% (43)	n/a
Science outcomes			
Biology/life science	9% (6)	7% (10)	n/a
Physical sciences	12% (8)	10% (15)	n/a
Other sciences	3% (2)	2% (3)	n/a
Mode of administration			
Standardized or state test	29% (20)	25% (37)	0%
Teacher-administered/grade	15% (10)	16% (23)	0%
Researcher-administered	60% (41)	59% (87)	0%
Timing of outcome measurement			
Mid-stream or immediate	99% (67)	88% (130)	0%
Follow-up	13% (9)	12% (17)	0%

Notes: Analysis based on 68 studies and 147 effect sizes. Study percentages can exceed 100% because a study may contribute effect sizes to more than one category (for example, researcher-developed and other measures). Personalization, interactions, grade level, and urbanicity categories are not mutually exclusive, so a study or effect size can appear in more than one row. *Mean (standard deviation) across effect sizes for continuous variables. Student race/ethnicity, SPED, EL, and economic disadvantage were reported almost only by U.S. studies. n/a = not applicable, because math and science subdomains apply only to outcomes in that content domain. PR = practice/rehearsal of information; DI = direct instruction; CL = collaborative learning; LD = learning through discovery; COI = conflict of interest; RDM = researcher-developed measure; SPED = special education; EL = English learner.

Table 2. Meta-regression results for the base model including study and method characteristics.

Moderator	Mean	*SE*	*m*	*k*	*df*	*p*
Peer review status					17.17	0.90
Peer reviewed	0.46	0.06	55	124	47.43	
Not peer reviewed	0.49	0.17	13	23	11.98	
Conflict of interest (COI) status					11.98	0.06
COI identified	1.01	0.28	10	18	8.93	
No COI identified	0.39	0.06	58	129	46.12	
Level of random assignment					67.00[a]	0.77
Students	0.45	0.08	42	92	35.60	
Teachers/classrooms	0.55	0.14	15	31	13.52	
Schools/centers	0.41	0.15	11	24	10.02	
Researcher-developed measure					13.93	0.93
Used RDM	0.47	0.07	49	105	39.72	
Did not use RDM	0.46	0.13	22	42	17.04	
Intervention focus					34.35	0.84
Math intervention	0.46	0.07	49	113	38.28	
Science intervention	0.49	0.14	19	34	20.48	

Notes: Mean = conditional mean, the predicted effect size at each level of the moderator with all other moderators held at their sample means. *p* tests whether effects differ across levels of the moderator, not whether a conditional mean differs from 0. For two-level moderators, *p* and *df* come from the robust (CR2) *t*-test with Satterthwaite degrees of freedom. [a]Robust Wald *F*-test of all category contrasts, using $m - 1$ denominator degrees of freedom, where *m* is the number of studies. Intervention focus and outcome domain are identical in this sample, so a single content domain indicator is used. *SE* = standard error; *m* = number of studies; *k* = number of effect sizes; *df* = degrees of freedom.

Table 3. Meta-regression results for a model including intervention characteristics, adjusting for study and method characteristics.

Moderator	Mean or *b*	*SE*	*m*	*k*	*df*	*p*
Intervention duration in weeks	0.05[b]	0.10	-	-	22.29	0.59
Intensity (hours per week)	0.19[b]	0.09	-	-	20.27	0.04
Frequency of Interaction					67.00[a]	0.47
Low (less than weekly)	0.31	0.26	5	10		
Moderate (1-2 times per week)	0.37	0.12	21	45		
High (3 or more time per week)	0.54	0.10	26	64		
Personalized instruction supports					67.00[a]	0.11
No personalization	0.19	0.15	13	24		
Provides feedback	0.53	0.07	49	113		
Adapts to learner aptitude/ability	0.41	0.14	24	46		
Instructional practices (digital app)					67.00[a]	0.19
PR only	0.22	0.11	20	53		
CL/LD	0.46	0.22	13	20		
PR + CL/ LD	0.36	0.20	6	17		
PR + CL/ LD + DI	0.52	0.25	8	12		
PR + DI	0.74	0.17	21	45		
Instructor facilitation					14.84	0.48
Not facilitated by instructor	0.49	0.08	50	99		
Facilitated by instructor	0.37	0.13	19	48		
Interactions within digital app					67.00[a]	0.14
No interactions	0.57	0.08	52	103		
Interactions with teachers	0.12	0.28	7	21		
Interactions with peers	0.27	0.16	14	37		
Intervention access					67.00[a]	0.60
Home	0.40	0.26	6	14		
School site, during school day	0.49	0.07	60	120		
School site, after school	0.16	0.32	3	13		

Notes: The model adjusts for study and method characteristics (Table 3). Missing moderator values were imputed (20 imputations) and estimates pooled with Rubin's rules. Mean = conditional mean, the predicted effect size at each level of the moderator with other moderators held at their sample means. Personalized instruction supports and interactions within the digital app are not mutually exclusive; the reference row sets both features to absent, and each feature row sets that feature to present with the other at its sample mean. *p* tests whether effects differ across levels of the moderator, not whether a conditional mean differs from 0. For two-level moderators and slopes, *p* and *df* come from the robust (CR2) *t*-test. [a]Robust Wald *F*-test of all category contrasts pooled across imputations, using $m - 1$ denominator degrees of freedom, where m is the number of studies. [b]Slope per one-unit increase in the natural log of weeks or hours per week. *SE* = standard error; *m* = number of studies; *k* = number of effect sizes; *df* = degrees of freedom; PR = practice/rehearsal of information; DI = direct instruction; CL = collaborative learning; LD = learning through discovery.

Table 4. Meta-regression results for sample characteristics (Panel A), setting characteristics (Panel B), and outcome characteristics (Panel C), each added to the base model.

Moderator	Mean or b	*SE*	*m*	*k*	*df*	*p*
Panel A. Sample characteristics						
Proportion female[b]	0.41	0.49	-	-	16.24	0.42
Grade level					67.00[a]	0.30
Elementary only (PreK-5)	0.56	0.09	31	73		
Middle (6-8)	0.33	0.10	17	31		
High (9-12)	0.34	0.11	24	48		
Panel B. Setting characteristics						
Country income level					67.00[a]	<0.01
Low or lower-middle income	0.76	0.19	12	26		
Upper-middle income	0.84	0.15	13	23		
High income	0.28	0.07	43	98		
Urbanicity					67.00[a]	0.99
Urban only	0.46	0.13	15	40		
Rural	0.43	0.18	14	33		
Suburban	0.46	0.22	8	11		
Panel C. Outcome characteristics						
Math subdomain (math outcomes)					67.00[a]	0.94
General or mixed math	0.50	0.12	20	43		
Number and operations	0.45	0.08	21	42		
Algebra, geometry, and data	0.46	0.11	14	28		
Science subdomain (science outcomes)					67.00[a]	0.07
Biology/life science	0.05	0.22	6	10		
Physical sciences	0.60	0.18	8	15		
General, earth, or other science	0.65	0.15	6	9		
Mode of measure administration					67.00[a]	0.92
Researcher-administered	0.45	0.09	41	87		
Standardized or state test	0.51	0.12	20	37		
Teacher-administered or course grade	0.44	0.11	10	23		
Timing of outcome measurement					7.35	0.81
Immediate	0.47	0.07	67	130		
Follow-up after intervention	0.43	0.13	9	17		

Notes. Each panel is a separate model that adds the listed characteristics to the base model of study and method characteristics (Table 2). Missing values of proportion female and urbanicity were imputed (20 imputations) and estimates pooled with Rubin's rules. Mean = conditional mean, the predicted effect size with other moderators held at their sample means. Grade levels and urbanicity categories are not mutually exclusive. The reference row is samples in elementary grades only or in urban settings only, and each other row sets that category to present with the others at their sample means. Country income level follows the World Bank FY2025 classification. Math and science subdomain means are computed for math and science outcomes, respectively. Immediate outcomes were measured during or directly after the intervention. *p* tests whether effects differ across levels of the moderator, not whether a conditional mean differs from 0. [a]Robust Wald *F* test pooled across imputations, using $m - 1$ denominator degrees of freedom, where *m* is the number of studies. [b]Slope per 1.0 change in the proportion; robust (CR2) *t* test. Exploratory U.S.-only results for other student characteristics are in Supplemental Material S3. *SE* = standard error; *m* = number of studies; *k* = number of effect sizes; *df* = degrees of freedom.

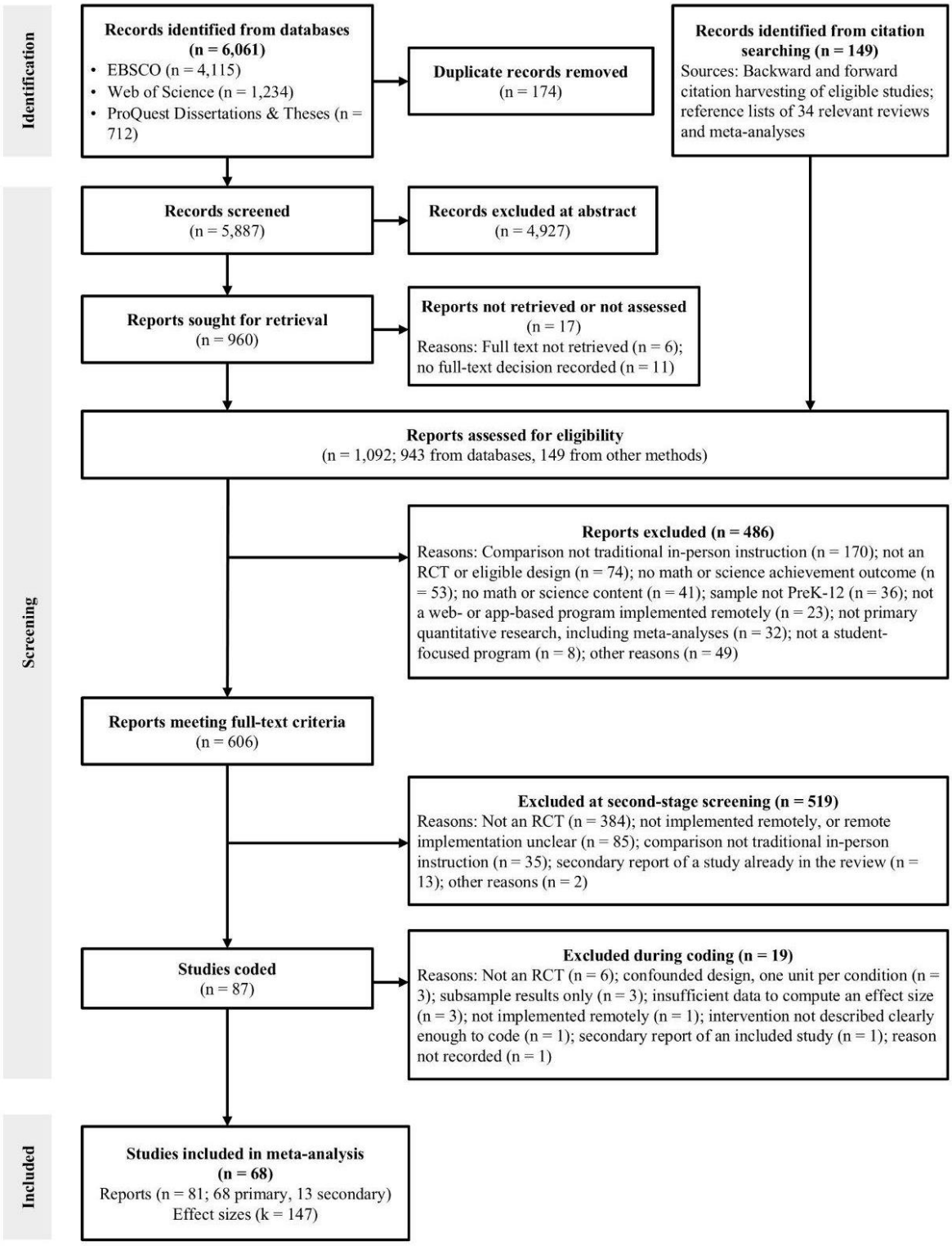


Figure 1. PRISMA 2020 Flow Diagram of Study Identification, Screening, and Inclusion

Supplemental Materials

S1. Database search strings
S2. Research organizations and review libraries searched
S3. Eligibility criteria
S4. Exploratory analysis of student demographic characteristics, U.S. studies only
S5. MetaForest procedure and results
S6. Sensitivity analyses, selective reporting bias, funnel plot, and forest plot
S7. Final meta-regression model
S8. Heterogeneity and variance explained across models
S9. References for included studies

The coding protocol is available on the Open Science Framework (https://osf.io/hf94y/overview?view_only=1aadfdd0161b4be1ac8fbda40502c3d6).

Supplemental Material S1. Database Search Strings

The search combined four concept blocks with AND. Each block was searched in titles (TI) and abstracts (AB). The strings below are those used in EBSCO and were adapted to the syntax of Web of Science Core Collection and ProQuest Dissertations & Theses Global. Searches were limited to English-language records published from January 2007 through December 2022.

Block 1. PreK–12 population

TI("Elementary education" OR "elementary school*" OR "elementary-aged" OR "elementary secondary" OR "primary education" OR "primary school*" OR "middle school*" OR "secondary school*" OR "secondary education" OR "secondary-aged" OR "high school*" OR highschool* OR "junior high" OR adolescen* OR teen* OR preteen* OR preadolescen* OR youth OR youths OR "young people" OR "pre-school*" OR preschool* OR "pre-kindergarten*" OR prekindergarten* OR preK* OR "pre-K" OR "pre-K12" OR "pre-K-12" OR "pre-K-12th" OR kindergarten* OR grade* OR classroom* OR K12 OR "K–12*" OR "PK–12*") OR AB("Elementary education" OR "elementary school*" OR "elementary-aged" OR "elementary secondary" OR "primary education" OR "primary school*" OR "middle school*" OR "secondary school*" OR "secondary education" OR "secondary-aged" OR "high school*" OR highschool* OR "junior high" OR adolescen* OR teen* OR preteen* OR preadolescen* OR youth OR youths OR "young people" OR "pre-school*" OR preschool* OR "pre-kindergarten*" OR prekindergarten* OR preK* OR "pre-K" OR "pre-K12" OR "pre-K-12" OR "pre-K-12th" OR kindergarten* OR grade* OR classroom* OR K12 OR "K–12*" OR "PK–12*") OR EL("early childhood" OR "Elementary education" OR "elementary secondary" OR grade* OR "high school*" OR "junior high" OR kindergarten* OR "middle school*" OR preschool* OR "primary education" OR "secondary education") OR TI child* OR AB child* OR SU child* OR SO child* OR TI school* OR AB school* OR SU school* OR SO school*

AND

Block 2. Digital and remote delivery

TI("distance education*" OR "distance learning" OR "distance instruction" OR "distance school*" OR "distance class*" OR "distance course*" OR "e-course*" OR "e-class*" OR "e-learning" OR "e-school*" OR "e-instruction" OR "e-education*" OR "virtual class*" OR "virtual course*" OR "virtual learning" OR "virtual school*" OR "virtual instruction" OR "virtual education*" OR "web-based education*" OR "web-based class*" OR "web-based school*" OR "web-based instruction" OR "web-based learning" OR "web-based course*" OR "remote education*" OR "remote class*" OR "remote course*" OR "remote learning" OR "remote school*" OR "remote instruction" OR "blended class*" OR "blended course*" OR "blended learning" OR "blended school*" OR "blended instruction" OR "blended education*" OR "hybrid education*" OR "hybrid class*" OR "hybrid course*" OR "hybrid learning" OR "hybrid school*" OR "hybrid instruction" OR "home-based education*" OR "home-based class*" OR "home-based course*" OR "home-based learning" OR "home-based school*" OR "home-based instruction" OR "digital class*" OR "digital learning" OR "digital instruction" OR "digital education*" OR "digital school*" OR "digital course*" OR "computer aided" OR "computer driven" OR "computer assisted" OR "online" OR "remote teaching" OR "educational technology") OR AB("distance education*" OR "distance learning" OR "distance instruction" OR "distance school*" OR "distance class*" OR "distance course*" OR "e-course*" OR "e-class*" OR "e-learning" OR "e-school*" OR "e-instruction" OR "e-education*" OR "virtual class*" OR "virtual course*" OR "virtual learning" OR "virtual school*" OR "virtual instruction" OR "virtual education*" OR "web-based education*" OR "web-based class*" OR "web-based school*" OR "web-based instruction" OR "web-based learning" OR "web-based course*" OR "remote education*" OR "remote class*" OR "remote course*" OR "remote learning" OR "remote school*" OR "remote instruction" OR "blended class*" OR "blended course*" OR "blended learning" OR "blended school*" OR "blended instruction" OR "blended education*" OR "hybrid education*" OR "hybrid class*" OR "hybrid course*" OR "hybrid learning" OR "hybrid school*" OR "hybrid instruction" OR "home-based education*" OR "home-based class*" OR "home-based course*" OR "home-based learning" OR "home-based school*" OR "home-based instruction" OR "digital class*" OR "digital learning" OR "digital instruction" OR "digital education*" OR "digital school*" OR "digital course*" OR "computer aided" OR "computer driven" OR "computer assisted" OR "online" OR "remote teaching" OR "educational technology")

AND

Block 3. Mathematics and science content

TI(Algebra* OR Arithmetic* OR Calculus OR Equation* OR Fraction* OR Geometr* OR "Integrated Math*" OR Math OR Mathematic* OR premathematic* OR "measurement task*" OR "area measurement" OR "measurement units" OR "measurement concept*" OR "measurement skills" OR "number skill*" OR "number sense" OR "number concept*" OR "number system*" OR "number processing" OR "number line" OR "number fact*" OR Numeracy OR Numerical OR "Pre-algebra*" OR Prealgebra* OR "probability concept*" OR "experimental probability" OR "theoretical probability" OR "probability unit*" OR "teaching probability" OR "teach probability" OR "Problem solving" OR "proof scheme*" OR "argument proof*" OR "proof process*" OR "proof theory" OR "proof writing" OR "ratio and proportion" OR "ratios and proportions" OR "ratio proportion" OR "ratios proportions" OR "proportional thinking" OR "proportional reasoning" OR "proportional structures" OR "fluid reasoning" OR "quantitative reasoning" OR "relational reasoning" OR "probabilistic reasoning" OR "reasoning strategies" OR "logical reasoning" OR "multiplicative reasoning" OR "analogical reasoning" OR Spatial* OR Trigonometr* OR decimal* OR "functions" OR "matrices" OR "place value" OR "probability and statistics" OR "statistics and probability" OR "science learning" OR "science knowledge" OR "scientific knowledge" OR "science achievement" OR "science understanding" OR "attitude* toward science" OR "science outcome*" OR "nature of science" OR "school science" OR "elementary science" OR "primary science" OR "secondary science" OR "science skill*") OR AB(Algebra* OR Arithmetic* OR Calculus OR Equation* OR Fraction* OR Geometr* OR "Integrated Math*" OR Math OR Mathematic* OR premathematic* OR "measurement task*" OR "area measurement" OR "measurement units" OR "measurement concept*" OR "measurement skills" OR "number skill*" OR "number sense" OR "number concept*" OR "number system*" OR "number processing" OR "number line" OR "number fact*" OR Numeracy OR Numerical OR "Pre-algebra*" OR Prealgebra* OR "probability concept*" OR "experimental probability" OR "theoretical probability" OR "probability unit*" OR "teaching probability" OR "teach probability" OR "Problem solving" OR "proof scheme*" OR "argument proof*" OR "proof process*" OR "proof theory" OR "proof writing" OR "ratio and proportion" OR "ratios and proportions" OR "ratio proportion" OR "ratios proportions" OR "proportional thinking" OR "proportional reasoning" OR "proportional structures" OR "fluid reasoning" OR "quantitative reasoning" OR "relational reasoning" OR "probabilistic reasoning" OR "reasoning strategies" OR "logical reasoning" OR "multiplicative reasoning" OR "analogical reasoning" OR Spatial* OR Trigonometr* OR decimal* OR "functions" OR "matrices" OR "place value" OR "probability and statistics" OR "statistics and probability" OR "science learning" OR "science knowledge" OR "scientific knowledge" OR "science achievement" OR "science understanding" OR "attitude* toward science" OR "science outcome*" OR "nature of science" OR "school science" OR "elementary science" OR "primary science" OR "secondary science" OR "science skill*")

AND

Block 4. Evaluation design

TI("Effectiveness stud*" OR "Effectiveness trial*" OR "Efficacy stud*" OR "Efficacy trial*" OR Evaluation* OR "Non-randomized" OR Nonrandomized OR QED* OR "Quasi-experiment*" OR Quasiexperiment* OR "Random* assign*" OR "Random* allocat*" OR Randomiz* OR Randomis* OR RCT* OR Trial OR Trials OR RDD* OR "regression discontinuity") OR AB("Effectiveness stud*" OR "Effectiveness trial*" OR "Efficacy stud*" OR "Efficacy trial*" OR Evaluation* OR "Non-randomized" OR Nonrandomized OR QED* OR "Quasi-experiment*" OR Quasiexperiment* OR "Random* assign*" OR "Random* allocat*" OR Randomiz* OR Randomis* OR RCT* OR Trial OR Trials OR RDD* OR "regression discontinuity")

Supplemental Material S2. Research Organizations and Review Libraries Searched

We manually searched the websites and publication databases of the following organizations for unpublished reports.

Abt Associates
American Institutes for Research
Brookings Institution
Campbell Collaboration (Campbell Systematic Reviews)
Center for Economic and Policy Research
Center for Science and Engineering Partnerships
EPPI-Centre
Institute for Policy Research
Institute for Policy Studies
Institute for Research on Poverty
Johns Hopkins Institute for Health and Social Policy
Mathematica
MDRC
Multidisciplinary Program in Inequality and Social Policy
National Bureau of Economic Research (NBER Working Papers)
NORC at the University of Chicago
Public Policy Associates
RAND Corporation
Ray Marshall Center
RTI International
Social Policy Research Associates
SRI International
WestEd

Supplemental Material S3. Eligibility Criteria

Table S3. Eligibility Criteria for This Review

	Criterion
1	The study sample included students in grades PreK–12.
2	The study evaluated a student-focused web- or app-based program providing mathematics or science instructional content.
3	The program delivered instruction without direct, in-person content delivery or continuous real-time guidance from a teacher. Students accessed the program outside school, in a different room from the instructor, or in the same room as an instructor whose only intended role was to help them access the program. Programs used to supplement teacher-led instruction, including blended models in which the teacher remained the primary source of content, were excluded.
4	The study randomly assigned participants to the program or to a comparison group that received traditional in-person classroom instruction without the program.
5	The study included a measure of students' achievement, academic readiness, knowledge, or skills in mathematics or science.
6	The study reported findings in a way that allowed an effect size to be computed, either directly or after contacting the authors.
7	The study was published between January 2007 and December 2022.
8	The study was written in or translated into English. Studies from any country were eligible.

Note. Criteria 1, 2, and 5 through 8 were applied during full-text screening under the review protocol. The protocol also allowed quasi-experimental designs and a broader range of comparison conditions. Criteria 3 and 4 reflect the second round of screening applied for this paper (see Methods and Figure 1).

Supplemental Material S4. Exploratory Meta-Regression Results for Student Demographic Characteristics, U.S. Studies Only

Table S4. Exploratory Meta-Regression Results for Student Demographic Characteristics, U.S. Studies Only

Characteristic	*b*	*SE*	*m*	*k*	*df*	*p*
Proportion Black	−0.38	0.31	11	22	1.74[a]	0.36
Proportion Hispanic	−0.33	0.38	11	26	2.30[a]	0.47
Proportion Asian	−9.73	1.50	8	16	3.14[a]	<0.01
Proportion White	0.83	0.28	9	18	3.10[a]	0.06
Proportion receiving special education	−0.45	0.21	7	22	1.28[a]	0.23
Proportion English learners	−0.55	0.32	7	16	1.91[a]	0.24
Proportion economically disadvantaged	−0.22	0.30	9	20	4.26	0.49

Note. These characteristics were reported almost only by U.S. studies, and most U.S. studies did not report them. Each row is a separate meta-regression of effect sizes on the characteristic, using only the U.S. studies that reported it, with no other moderators. *b* = change in effect size per 1.0 change in the proportion, with robust (CR2) standard errors and Satterthwaite degrees of freedom. The proportion of Asian students had a very narrow range ($M = 0.05$, $SD = 0.03$), so its slope implies a change of about −0.30 per standard deviation. [a]Fewer than 4 degrees of freedom; the test is unreliable and the estimate should be read as descriptive. Results are exploratory. *SE* = standard error; *m* = number of studies; *k* = number of effect sizes; *df* = degrees of freedom.

Supplemental Material S5. MetaForest Procedure and Results

We used MetaForest (Van Lissa, 2020) as an exploratory check on the meta-regression results. MetaForest adapts random forests to meta-analysis by weighting effect sizes, here with random-effects weights, and by accounting for the clustering of effect sizes within studies. It does not assume linear or additive relationships and ranks moderators by permutation importance, the loss in predictive accuracy when a moderator's values are randomly shuffled. It does not test moderators for statistical significance.

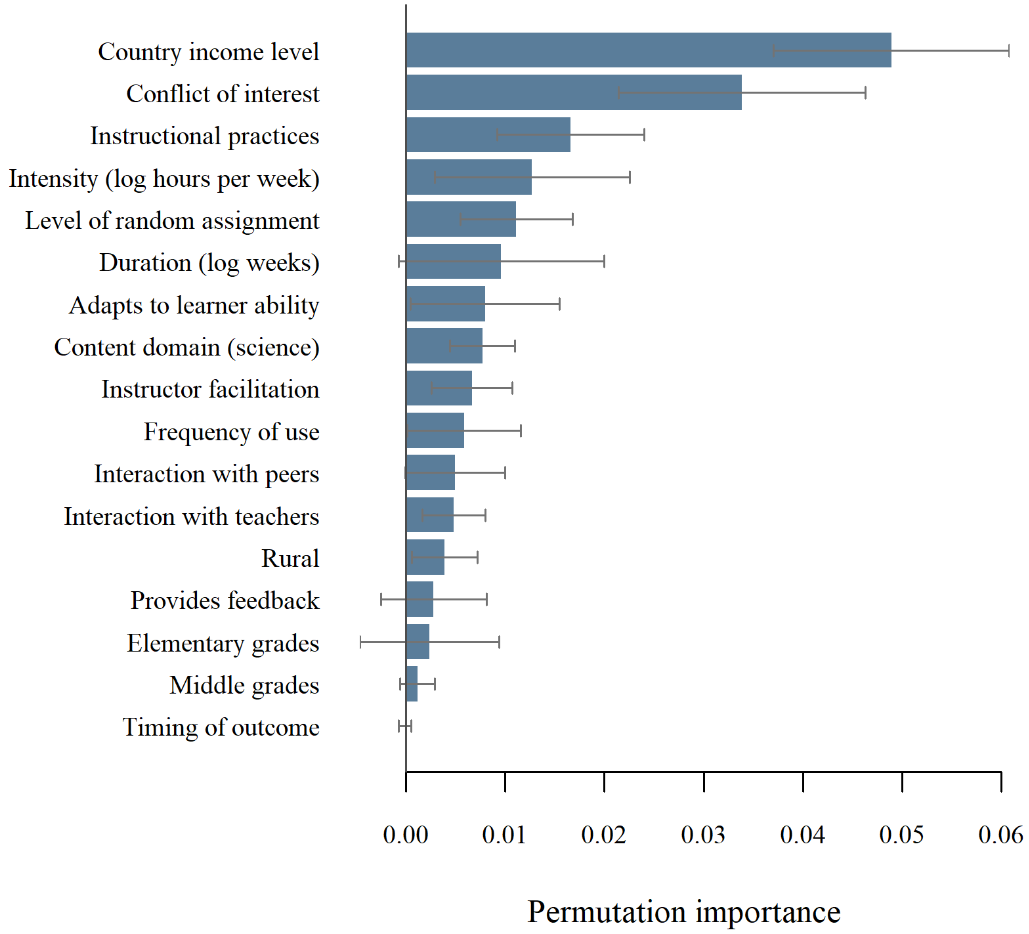
We applied MetaForest to each of the 20 imputed datasets, combining the imputed values from all moderator blocks. Student demographic characteristics other than gender were excluded because of missing data. Each categorical moderator entered as a single variable. To reduce the number of moderators, we used recursive preselection with 2,000 trees and 10 replications per imputed dataset, for 200 replications in total, and kept moderators selected in at least half of them. We then fit a final forest of 5,000 trees on the retained moderators in each imputed dataset and averaged variable importance across datasets.

Preselection retained 17 of 24 candidate moderators (Table S5). Country income level was retained in every replication and had the highest importance, followed by conflict of interest, instructional practices, intensity, and level of random assignment (Figure S1). The final forests had a mean out-of-bag R^2 of .12. Two results differed from the meta-regressions. Adaptivity ranked above feedback, likely because most adaptive programs also gave feedback, so the two features carry overlapping information. Outcome subdomain was not retained, likely because it combined mathematics subdomains, which showed no differences, with the science subdomains that did.

Table S5. MetaForest Preselection and Variable Importance

Moderator	% of replications retained	Retained	Importance	Rank
Country income level	100.0	Yes	0.0489	1
Conflict of interest	96.5	Yes	0.0338	2
Instructional practices	94.5	Yes	0.0166	3
Intensity (log hours per week)	60.5	Yes	0.0127	4
Level of random assignment	64.5	Yes	0.0111	5
Duration (log weeks)	55.5	Yes	0.0096	6
Adapts to learner ability	70.5	Yes	0.0080	7
Content domain (science)	90.0	Yes	0.0077	8
Instructor facilitation	88.0	Yes	0.0067	9
Frequency of use	67.0	Yes	0.0059	10
Interaction with peers	66.0	Yes	0.0050	11
Interaction with teachers	92.5	Yes	0.0048	12
Rural	54.0	Yes	0.0039	13
Provides feedback	57.5	Yes	0.0028	14
Elementary grades	75.5	Yes	0.0024	15
Middle grades	61.5	Yes	0.0012	16
Timing of outcome	69.5	Yes	-0.0001	17
Suburban	48.5	No	–	–
Outcome subdomain	38.0	No	–	–
High school grades	17.5	No	–	–
Access setting	14.0	No	–	–
Measure administration	12.0	No	–	–
Researcher-developed measure	10.5	No	–	–
Peer review status	1.5	No	–	–

Note. Retention is the percentage of 200 preselection replications (10 per imputed dataset) in which the moderator was kept. Moderators retained in at least 50% of replications entered the final forests. Importance is mean permutation importance across the 20 imputed datasets. Negative values indicate no predictive value.



Note. Bars show mean permutation importance across the 20 imputed datasets, and error bars show one standard deviation across datasets. Only moderators retained by preselection are shown.

Figure S5. Variable Importance From the Final MetaForest Models

Supplemental Material S6. Sensitivity analyses, selective reporting bias, funnel plot, and forest plot

Table S6. Sensitivity of the Overall Average Effect to Modeling Choices, Influential Studies, and Selective Reporting

Analysis	g	*SE*	95% CI	τ
Main model (ρ = 0.50)	0.47	0.06	[0.35, 0.60]	0.52
Assumed within-study correlation				
ρ = 0.00	0.48	0.07	[0.35, 0.61]	0.51
ρ = 0.20	0.48	0.07	[0.35, 0.61]	0.51
ρ = 0.80	0.47	0.06	[0.34, 0.60]	0.52
Resampling and influence				
Cluster bootstrap	0.47	–	[0.34, 0.63]	0.52
Leave one study out	0.44 to 0.49	0.06 to 0.07	–	0.46 to 0.53
Influential studies removed	0.44 to 0.46	0.06	–	0.46 to 0.50
Selective reporting				
Three-parameter selection model	0.49	0.10	[0.31, 0.70]	0.48
Trim and fill	0.40	0.04	[0.32, 0.49]	–

Note. The main model is the random-effects model with no moderators (68 studies, 147 effect sizes), using correlated and hierarchical effects with robust (CR2) standard errors. ρ = assumed correlation among effect sizes from the same sample; τ = total heterogeneity standard deviation. [a]1,000 resamples of studies with bias-corrected and accelerated (BCa) intervals. [b]Range across 68 models, each omitting one study; all estimates had $p < .001$. [c]Four studies had Cook's distance above $4/m$; the range shows estimates when each was removed one at a time. [d]One cut point at one-tailed $p = .025$, estimated with the metaselection R package, with 1,000 cluster bootstrap percentile intervals. The estimated relative probability of reporting a nonsignificant result was 1.53, 95% CI [0.70, 3.17]. [e]Univariate random-effects model that ignores dependence among effect sizes, reported for description only; no studies were imputed (see Figure S2). *SE* = standard error; CI = confidence interval.

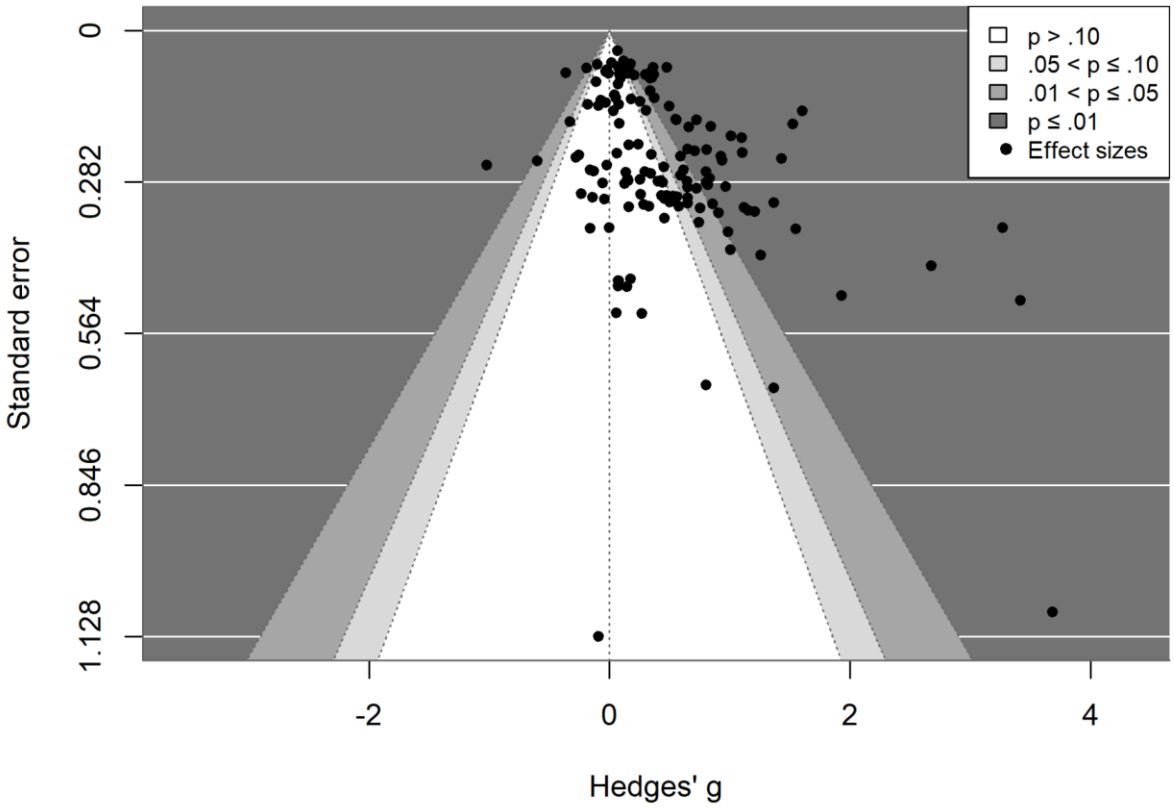


Figure S6-1. Contour-Enhanced Funnel Plot of Effect Sizes

Note. Each point is one of the 147 effect sizes, plotted against its standard error. The dotted line marks zero, and shaded regions show where an effect size would be statistically significant at the levels shown. The plot ignores dependence among effect sizes from the same study and is shown for description only (see Table S6 for formal tests).

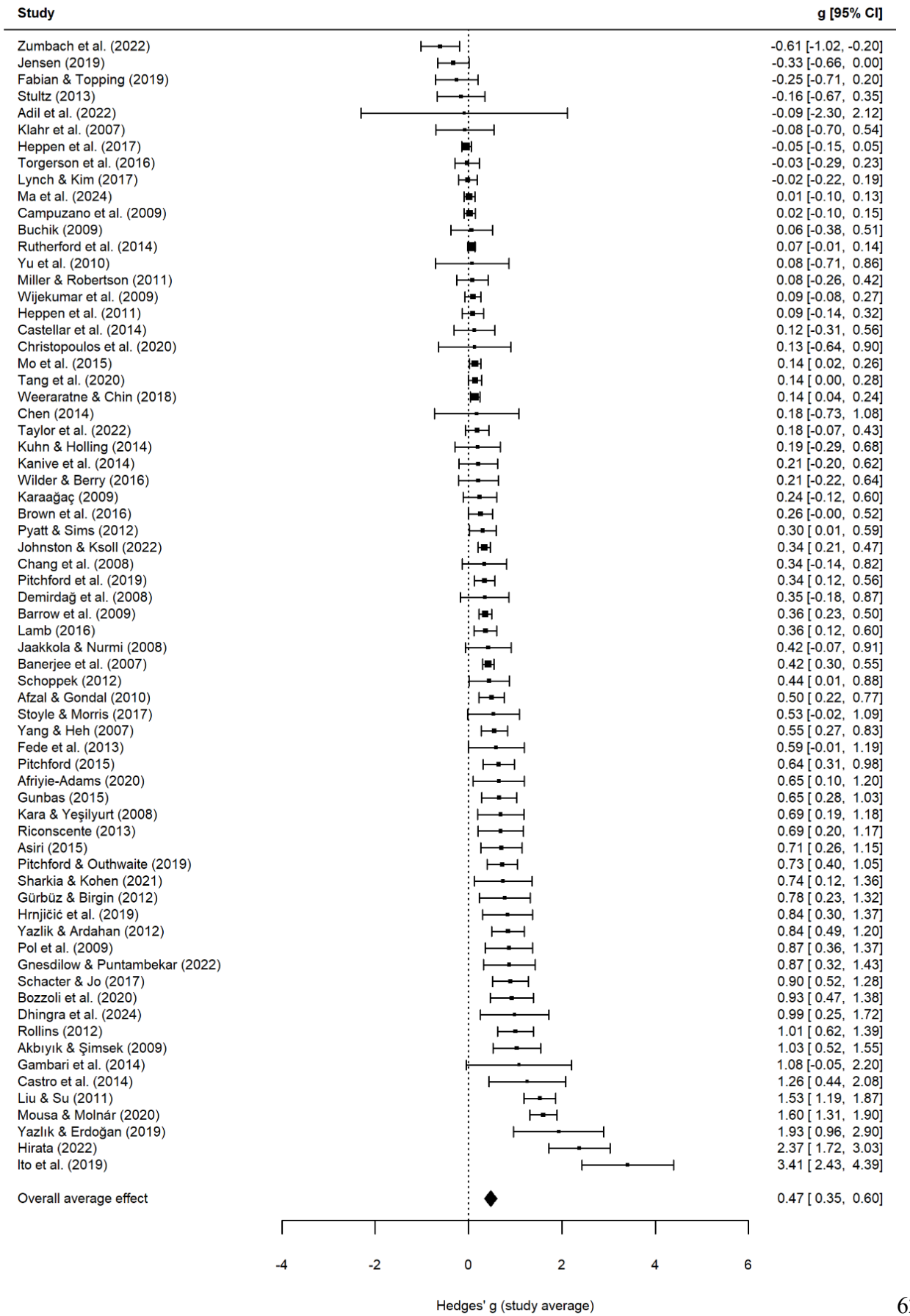

Study
g [95% CI]
Zumbach et al. (2022) -0.61 [-1.02, -0.20]
Jensen (2019) -0.33 [-0.66, 0.00]
Fabian & Topping (2019) -0.25 [-0.71, 0.20]
Stultz (2013) -0.16 [-0.67, 0.35]
Adil et al. (2022) -0.09 [-2.30, 2.12]
Klahr et al. (2007) -0.08 [-0.70, 0.54]
Heppen et al. (2017) -0.05 [-0.15, 0.05]
Torgerson et al. (2016) -0.03 [-0.29, 0.23]
Lynch & Kim (2017) -0.02 [-0.22, 0.19]
Ma et al. (2024) 0.01 [-0.10, 0.13]
Campuzano et al. (2009) 0.02 [-0.10, 0.15]
Buchik (2009) 0.06 [-0.38, 0.51]
Rutherford et al. (2014) 0.07 [-0.01, 0.14]
Yu et al. (2010) 0.08 [-0.71, 0.86]
Miller & Robertson (2011) 0.08 [-0.26, 0.42]
Wijekumar et al. (2009) 0.09 [-0.08, 0.27]
Heppen et al. (2011) 0.09 [-0.14, 0.32]
Castellar et al. (2014) 0.12 [-0.31, 0.56]
Christopoulos et al. (2020) 0.13 [-0.64, 0.90]
Mo et al. (2015) 0.14 [0.02, 0.26]
Tang et al. (2020) 0.14 [0.00, 0.28]
Weeraratne & Chin (2018) 0.14 [0.04, 0.24]
Chen (2014) 0.18 [-0.73, 1.08]
Taylor et al. (2022) 0.18 [-0.07, 0.43]
Kuhn & Holling (2014) 0.19 [-0.29, 0.68]
Kanive et al. (2014) 0.21 [-0.20, 0.62]
Wilder & Berry (2016) 0.21 [-0.22, 0.64]
Karaağaç (2009) 0.24 [-0.12, 0.60]
Brown et al. (2016) 0.26 [-0.00, 0.52]
Pyatt & Sims (2012) 0.30 [0.01, 0.59]
Johnston & Ksoll (2022) 0.34 [0.21, 0.47]
Chang et al. (2008) 0.34 [-0.14, 0.82]
Pitchford et al. (2019) 0.34 [0.12, 0.56]
Demirdağ et al. (2008) 0.35 [-0.18, 0.87]
Barrow et al. (2009) 0.36 [0.23, 0.50]
Lamb (2016) 0.36 [0.12, 0.60]
Jaakkola & Nurmi (2008) 0.42 [-0.07, 0.91]
Banerjee et al. (2007) 0.42 [0.30, 0.55]
Schoppek (2012) 0.44 [0.01, 0.88]
Afzal & Gondal (2010) 0.50 [0.22, 0.77]
Stoyle & Morris (2017) 0.53 [-0.02, 1.09]
Yang & Heh (2007) 0.55 [0.27, 0.83]
Fede et al. (2013) 0.59 [-0.01, 1.19]
Pitchford (2015) 0.64 [0.31, 0.98]
Afriyie-Adams (2020) 0.65 [0.10, 1.20]
Gunbas (2015) 0.65 [0.28, 1.03]
Kara & Yeşilyurt (2008) 0.69 [0.19, 1.18]
Riconscente (2013) 0.69 [0.20, 1.17]
Asiri (2015) 0.71 [0.26, 1.15]
Pitchford & Outhwaite (2019) 0.73 [0.40, 1.05]
Sharkia & Kohen (2021) 0.74 [0.12, 1.36]
Gürbüz & Birgin (2012) 0.78 [0.23, 1.32]
Hrnjičić et al. (2019) 0.84 [0.30, 1.37]
Yazlik & Ardahan (2012) 0.84 [0.49, 1.20]
Pol et al. (2009) 0.87 [0.36, 1.37]
Gnesdilow & Puntambekar (2022) 0.87 [0.32, 1.43]
Schacter & Jo (2017) 0.90 [0.52, 1.28]
Bozzoli et al. (2020) 0.93 [0.47, 1.38]
Dhingra et al. (2024) 0.99 [0.25, 1.72]
Rollins (2012) 1.01 [0.62, 1.39]
Akbıyık & Şimsek (2009) 1.03 [0.52, 1.55]
Gambari et al. (2014) 1.08 [-0.05, 2.20]
Castro et al. (2014) 1.26 [0.44, 2.08]
Liu & Su (2011) 1.53 [1.19, 1.87]
Mousa & Molnár (2020) 1.60 [1.31, 1.90]
Yazlık & Erdoğan (2019) 1.93 [0.96, 2.90]
Hirata (2022) 2.37 [1.72, 3.03]
Ito et al. (2019) 3.41 [2.43, 4.39]
Overall average effect 0.47 [0.35, 0.60]
-4
-2
0
2
4
6
Hedges' g (study average)

Figure S6-2. Study-Level Effect Sizes

Note. Each row shows one study's average effect size (Hedges' *g*) and 95% confidence interval, combining the study's effect sizes under an assumed within-study correlation of 0.50. Studies are ordered by effect size. The diamond shows the overall average effect from the main random-effects model ($g = 0.47$, 95% CI [0.35, 0.60]). Full references are in Supplemental Material S9.

Supplemental Material S7: Final Meta-Regression Model

Table S7. Final Meta-Regression Model

Moderator	Mean or b	*SE*	*m*	*k*	*df*	*p*
Study and method characteristics						
Peer review status					17.09	0.52
Peer reviewed	0.44	0.06	55	124		
Not peer reviewed	0.52	0.12	13	23		
Conflict of interest (COI)					11.62	0.06
COI identified	0.88	0.22	10	18		
No COI identified	0.39	0.05	58	129		
Level of random assignment					67.00[a]	0.048
Students	0.41	0.06	42	92		
Teachers/classrooms	0.69	0.12	15	31		
Schools/centers	0.32	0.10	11	24		
Researcher-developed measure (RDM)					15.47	0.43
Used RDM	0.48	0.06	49	105		
Did not use RDM	0.37	0.11	22	42		
Content domain					9.61	0.20
Math	0.50	0.07	49	113		
Science	0.28	0.13	19	34		
Program characteristics						
Intensity (log hours per week)	0.05[b]	0.06	–	–	17.32	0.40
Personalized instruction supports					67.00[a]	0.08
No personalization	0.14	0.17	13	24		
Provides feedback	0.51	0.06	49	113		
Adapts to learner ability	0.62	0.10	24	46		
Instructional practices					67.00[a]	0.27
PR only	0.40	0.07	20	53		
CL/LD only	0.39	0.13	13	20		
PR + CL/LD	0.21	0.14	6	17		
PR + CL/LD + DI	0.45	0.25	8	12		
PR + DI	0.62	0.12	21	45		
Outcome characteristics						
Science subdomain (science outcomes)					67.00[a]	<0.01
Biology/life science	−0.02	0.12	6	10		
Physical sciences	0.69	0.20	8	15		
General, earth, or other science	0.61	0.14	6	9		
Setting characteristics						
Country income level					67.00[a]	<0.01
Low or lower-middle income	0.87	0.19	12	26		
Upper-middle income	0.86	0.15	13	23		
High income	0.24	0.08	43	98		

Note. All moderators were entered in one model. Moderators were included if their overall test in the block model had $p < .10$ or at least one contrast had $p < .05$. Missing values were imputed (20 imputations) and estimates pooled with Rubin's rules. Mean = conditional mean, the predicted effect size with other moderators held at their sample

means. Personalization categories are not mutually exclusive; the reference row is programs with neither feature, and each other row sets that feature to present with the other at its sample mean. Science subdomain means are computed for science outcomes. Because the model includes science subdomain indicators, the content domain means are not a contrast between content areas (see Table 2). *p* tests whether effects differ across levels of the moderator. [a]Robust Wald *F* test pooled across imputations, using $m - 1$ denominator degrees of freedom. [b]Slope per 1-unit change in log hours per week; robust (CR2) *t* test. PR = practice or rehearsal; CL = collaborative learning; LD = learning through discovery; DI = direct instruction; *SE* = standard error; *m* = number of studies; *k* = number of effect sizes; *df* = degrees of freedom.

Supplemental Materials S8: Heterogeneity and variance explained across models

Table S8. Heterogeneity and Variance Explained Across Models

Model	τ	τ_B	τ_W	Reduction vs. unconditional	Reduction vs. base
Unconditional model	0.518	0.430	0.288	–	–
Base model (study and method)	0.507	0.416	0.290	2.1%	–
Base + outcome	0.505	0.407	0.299	2.3%	0.2%
Base + sample	0.501	0.408	0.291	3.1%	1.1%
Base + program	0.465	0.361	0.293	10.1%	8.2%
Base + setting	0.453	0.353	0.283	12.6%	10.7%
Final model (Table S7)	0.389	0.257	0.292	24.8%	23.2%

Note. Models are ordered by reduction in total heterogeneity. τ = total heterogeneity, the square root of the sum of the between-study and within-study variance components; τ_B = between-study standard deviation; τ_W = within-study standard deviation. Reduction = (τ of the comparison model − τ of the model) / τ of the comparison model. Estimates for models with imputed moderators are pooled across 20 imputations.

Supplemental Material S9. References for Included Studies

Adil, F., Nazir, R., & Akhtar, M. (2022). Investigating the impact on learning outcomes through the use of EdTech during COVID-19: Evidence from an RCT in the Punjab province of Pakistan. Frontiers in Education, 7, 993265. https://doi.org/10.3389/feduc.2022.993265

Afriyie-Adams, M. (2020). MathConceptz: An experimental design to improve math performance. Doctoral dissertation, Rutgers the State University of New Jersey, School of Graduate Studies.

Afzal, M. T., & Gondal, M. B. (2010). Effect of mathematics software facilitated teaching on students learning. International Journal of Technology, Knowledge and Society, 6(3), 111.

Akbıyık, C., & Şimsek, N. (2009). Accelerated learning in classroom and computer environments. Eurasian Journal of Educational Research, 37, 32-52.

Asiri, M. (2015). The effect of computer-assisted instruction on the students acquisition of science processes skills in chemistry course. Journal of Research in Curriculum Instruction and Educational Technology, 1(3), 61-73. https://journals.ekb.eg/article_24547.html

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. The Quarterly Journal of Economics, 122(3), 1235-1264. https://doi.org/10.1162/qjec.122.3.1235

Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. American Economic Journal: Economic Policy, 1(1), 52-74. https://www.aeaweb.org/articles?id=10.1257/pol.1.1.52

Bozzoli, C. G., Gonzalez-Gadea, M. L., Hermida, M. J., Navarro, L., Olego, T. A., & Klingberg, T. (2020). Digital, mathematical and cognitive training: Evidence from a randomized trial. PsyArXiv.

Brown, S. W., Lawless, K. A., Rhoads, C., Newton, S. D., & Lynn, L. (2016). Increasing students' science writing skills through a PBL simulation. International Association for Development of the Information Society. https://eric.ed.gov/?id=ED571428

Buchik, J. C. (2009). Increasing student learning of mathematics facts through online computer-assisted instruction. University of Houston-Clear Lake.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of reading and mathematics software products: Findings from two student cohorts. NCEE 2009-4041. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED504657

Castellar, E. N., Van Looy, J., Szmalec, A., & De Marez, L. (2014). Improving arithmetic skills through gameplay: Assessment of the effectiveness of an educational game in terms of cognitive and affective learning outcomes. Information sciences, 264, 19-31. https://doi.org/10.1016/j.ins.2013.09.030

Castro, M. V. D., Bissaco, M. A. S., Panccioni, B. M., Rodrigues, S. C. M., & Domingues, A. M. (2014). Effect of a virtual environment on the development of mathematical skills in children with dyscalculia. PloS one, 9(7), e103354. https://doi.org/10.1371/journal.pone.0103354

Chang, K. E., Sung, Y. T., Chen, Y. L., & Huang, L. H. (2008). Learning multiplication through computer-assisted learning activities. Computers in Human Behavior, 24(6), 2904-2916. https://doi.org/10.1016/j.chb.2008.04.015

Chen, M. (2014). Primary children's attitudes towards electronic games and effect of electronic games on primary children's mathematics learning. Doctoral dissertation, University of Dundee.

Christopoulos, A., Kajasilta, H., Salakoski, T., & Laakso, M. J. (2020). Limits and virtues of educational technology in elementary school mathematics. Journal of Educational Technology Systems, 49(1), 59-81. https://doi.org/10.1177/0047239520908838

Demirdağ, B., Kartal, M., & Tüysüz, C. (2008). Developing a computer assisted education material related to thermochemistry. Journal of Turkish Science Education, 5(3).

Dhingra, K., Aggarwal, R., Garg, A., Pujari, J., & Yadav, D. (2024). Mathlete: An adaptive assistive technology tool for children with dyscalculia. Disability and Rehabilitation: Assistive Technology, 19(1), 9-15. https://doi.org/10.1080/17483107.2022.2134473

Fabian, K., & Topping, K. J. (2019). Putting "mobile" into mathematics: Results of a randomised controlled trial. Contemporary Educational Psychology, 59, 101783. https://doi.org/10.1016/j.cedpsych.2019.101783

Fede, J. L., Pierce, M. E., Matthews, W. J., & Wells, C. S. (2013). The effects of a computer-assisted, schema-based instruction intervention on word problem-solving skills of low-performing fifth grade students. Journal of Special Education Technology, 28(1), 9-21. https://doi.org/10.1177/016264341302800102

Gambari, I. A., Ezenwa, V. I., & Anyanwu, R. C. (2014). Comparative effects of two modes of computer-assisted instructional packages on solid geometry achievement. Contemporary Educational Technology, 5(2), 110-120. https://dergipark.org.tr/en/pub/cet/issue/25736/271504

Gnesdilow, D., & Puntambekar, S. (2022). Comparing middle school students' science explanations during physical and virtual laboratories. Journal of Science Education and Technology, 31(2), 191-202. https://doi.org/10.1007/s10956-021-09941-0

Gunbas, N. (2015). Students' mathematics word problem-solving achievement in a computer-based story. Journal of Computer Assisted Learning, 31(1), 78-95. https://doi.org/10.1111/jcal.12067

Gürbüz, R., & Birgin, O. (2012). The effect of computer-assisted teaching on remedying misconceptions: The case of the subject "probability". Computers & Education, 58(3), 931-941. https://doi.org/10.1016/j.compedu.2011.11.005

Heppen, J. B., Sorensen, N., Allensworth, E., Walters, K., Rickles, J., Taylor, S. S., & Michelman, V. (2017). The struggle to pass algebra: Online vs. face-to-face credit recovery for at-risk urban students. Journal of Research on Educational Effectiveness, 10(2), 272-296. https://doi.org/10.1080/19345747.2016.1168500

Heppen, J. B., Walters, K., Clements, M., Faria, A. M., Tobey, C., Sorensen, N., & Culp, K. (2011). Access to Algebra I: The effects of online mathematics for grade 8 students. NCEE 2012-4021. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED527394

Hirata, G. (2022). Play to Learn: The impact of technology on students' math performance. Journal of Human Capital, 16(3), 437-459. https://doi.org/10.1086/719846

Hrnjičić, A., Buzađija, N., & Bikić, N. (2019). Efekti online digitalnog multimedijalnog sadržaja na učenička postignuća iz oblasti kvadratna funkcija i njen grafik (The effects of online digital multimedia content on student achievements in the area of quadratic function and its graph). Educa-Časopis za obrazovanje, nauku i kulturu.

Ito, H., Kasai, K., & Nakamuro, M. (2019). Does computer-aided instruction improve children's cognitive and non-cognitive skills?: Evidence from Cambodia. RIETI.

Jaakkola, T., & Nurmi, S. (2008). Fostering elementary school students' understanding of simple electricity by combining simulation and laboratory activities. Journal of Computer Assisted Learning, 24(4), 271-283. https://doi.org/10.1111/j.1365-2729.2007.00259.x

Jensen, D. C. (2019). A Quasi-Experimental Examination of Digital Literacy and Achievement of Learning Objectives in Online versus Traditional Labs. Doctoral dissertation, Northcentral University.

Johnston, J., & Ksoll, C. (2022). Effectiveness of interactive satellite-transmitted instruction: Experimental evidence from Ghanaian primary schools. Economics of Education Review, 91, 102315. https://doi.org/10.1016/j.econedurev.2022.102315

Kanive, R., Nelson, P. M., Burns, M. K., & Ysseldyke, J. (2014). Comparison of the effects of computer-based practice and conceptual understanding interventions on mathematics fact retention and generalization. The Journal of Educational Research, 107(2), 83-89. https://doi.org/10.1080/00220671.2012.759405

Kara, Y., & Yeşilyurt, S. (2008). Comparing the impacts of tutorial and edutainment software programs on students' achievements, misconceptions, and attitudes towards biology. Journal of Science Education and Technology, 17(1), 32-41. https://doi.org/10.1007/s10956-007-9077-z

Karaağaç, S. (2009). Genetik Bilimi'Nin Lise öğrencileri tarafından öğrenilmesinde simülasyonun rolü (The role of simulation in high school students' learning of genetic science). Master's thesis, Marmara Universitesi (Turkey).

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. Journal of Research in Science teaching, 44(1), 183-203. https://doi.org/10.1002/tea.20152

Kuhn, J. T., & Holling, H. (2014). Number sense or working memory? The effect of two computer-based trainings on mathematical skills in elementary school. Advances in cognitive psychology, 10(2), 59. https://doi.org/10.5709/acp-0157-2

Lamb, R. L. (2016). Examination of the effects of dimensionality on cognitive processing in science: A computational modeling experiment comparing online laboratory simulations and serious educational games. Journal of Science Education and Technology, 25(1), 1-15. https://doi.org/10.1007/s10956-015-9587-z

Liu, H. C., & Su, I. H. (2011). Learning residential electrical wiring through computer simulation: The impact of computer-based learning environments on student achievement and cognitive load. British Journal of Educational Technology, 42(4), 598-607. https://doi.org/10.1111/j.1467-8535.2009.01047.x

Lynch, K., & Kim, J. S. (2017). Effects of a summer mathematics intervention for low-income children: A randomized experiment. Educational Evaluation and Policy Analysis, 39(1), 31-53. https://doi.org/10.3102/0162373716662339

Ma, Y., Zhang, X., Abbey, C., Hu, D., Lee, O., Hung, W., Chang, C., Wu, C., Friesen, D. & Rozelle, S. (2024). Computer assisted learning and academic performance in rural Taiwan. Journal of Research on Educational Effectiveness, 17(4), 1056-1079. https://doi.org/10.1080/19345747.2023.2279167

Miller, D. J., & Robertson, D. P. (2011). Educational benefits of using game consoles in a primary classroom: A randomised controlled trial. British Journal of Educational Technology, 42(5), 850-864. https://doi.org/10.1111/j.1467-8535.2010.01114.x

Mo, D., Zhang, L., Wang, J., Huang, W., Shi, Y., Boswell, M., & Rozelle, S. (2015). Persistence of learning gains from computer assisted learning: Experimental evidence from China. Journal of Computer Assisted Learning, 31(6), 562-581. https://doi.org/10.1111/jcal.12106

Mousa, M., & Molnár, G. (2020). Computer-based training in math improves inductive reasoning of 9-to 11-year-old children. Thinking Skills and Creativity, 37, 100687. https://doi.org/10.1016/j.tsc.2020.100687

Pitchford, N. J. (2015). Development of early mathematical skills with a tablet intervention: a randomized control trial in Malawi. Frontiers in psychology, 6, 124709. https://doi.org/10.3389/fpsyg.2015.00485

Pitchford, N. J., Chigeda, A., & Hubber, P. J. (2019). Interactive apps prevent gender discrepancies in early-grade mathematics in a low-income country in sub-Sahara Africa. Developmental Science, 22(5), e12864. https://doi.org/10.1111/desc.12864

Pitchford, N. J., & Outhwaite, L. A. (2019). Secondary benefits to attentional processing through intervention with an interactive maths app. Frontiers in Psychology, 10, 2633. https://doi.org/10.3389/fpsyg.2019.02633

Pol, H. J., Harskamp, E. G., Suhre, C. J., & Goedhart, M. J. (2009). How indirect supportive digital help during and after solving physics problems can improve problem-solving abilities. Computers & Education, 53(1), 34-50. https://doi.org/10.1016/j.compedu.2008.12.015

Pyatt, K., & Sims, R. (2012). Virtual and physical experimentation in inquiry-based science labs: Attitudes, performance and access. Journal of Science Education and Technology, 21(1), 133-147. https://doi.org/10.1007/s10956-011-9291-6

Riconscente, M. M. (2013). Results from a controlled study of the iPad fractions game Motion Math. Games and Culture, 8(4), 186-214. https://doi.org/10.1177/1555412013496894

Rollins, K. F. (2012). The Effects of a Computerized Study Program on the Acquisition of Science Vocabulary. Doctoral dissertation, Baylor University.

Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., Richland, L., Tran, N., Schneider, S., Duran, L., & Martinez, M. E. (2014). A randomized trial of an elementary school mathematics software intervention: Spatial-temporal math. Journal of Research on Educational Effectiveness, 7(4), 358-383. https://doi.org/10.1080/19345747.2013.856978

Schacter, J., & Jo, B. (2017). Improving preschoolers' mathematics achievement with tablets: A randomized controlled trial. Mathematics Education Research Journal, 29(3), 313-327. https://doi.org/10.1007/s13394-017-0203-9

Schoppek, W. (2012). Dieser Beitrag wurde unter der geschäftsführenden Herausgeberschaft von Jens Möller angenommen (Dynamic task selection in learning arithmetic: The role of learner control and adaptation based on a hierarchy of skills). Zeitschrift für pädagogische Psychologie. https://doi.org/10.1024/1010-0652/a000059

Sharkia, H., & Kohen, Z. (2021). Flipped classroom among minorities in the context of mathematics learning: The Israeli case. Mathematics, 9(13), 1500. https://doi.org/10.3390/math9131500

Stoyle, K. L., & Morris, B. J. (2017). Blogging mathematics: Using technology to support mathematical explanations for learning fractions. Computers & Education, 111, 114-127. https://doi.org/10.1016/j.compedu.2017.04.007

Stultz, S. L. (2013). The effectiveness of computer-assisted instruction for teaching mathematics to students with specific learning disability. The Journal of Special Education Apprenticeship, 2(2), 7. https://doi.org/10.58729/2167-3454.1028

Tang, B., Ting, T. T., Wu, C. I., Ma, Y., Mo, D., Hung, W. T., & Rozelle, S. (2020). The impact of online computer assisted learning at home for disadvantaged children in Taiwan: Evidence from a randomized experiment. Sustainability, 12(23), 10092. https://doi.org/10.3390/su122310092

Taylor, J. A., Adams, C. T., Westbrook, A. L., Creasap Gee, J., Spybrook, J. K., Kowalski, S. M., Gardner, A. L., & Bloom, M. (2022). The effect of a student–teacher–scientist partnership program on high school students' science achievement and attitudes about scientists. Journal of Research in Science Teaching, 59(3), 423-457. https://doi.org/10.1002/tea.21733

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N. & Torgerson, D. (2016). Affordable Online Maths Tuition: Evaluation report and executive summary. Education Endowment Foundation. https://eric.ed.gov/?id=ED581116

Weeraratne, B., & Chin, B. (2018). Can Khan Academy e-learning video tutorials improve mathematics achievement in Sri Lanka?. International Journal of Education and Development using Information and Communication Technology, 14(3), 93-112. https://eric.ed.gov/?id=EJ1201489

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P., & Peck, K. (2009). A multisite cluster randomized trial of the effects of CompassLearning Odyssey [R] Math on the math achievement of selected grade 4 students in the Mid-Atlantic region. Final Report. NCEE 2009-4068. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED507314

Wilder, S., & Berry, L. (2016). Emporium Model: The key to content retention in secondary math courses. Journal of Educators Online, 13(2), 53-75. https://eric.ed.gov/?id=EJ1106734

Yang, K. Y., & Heh, J. S. (2007). The impact of internet virtual physics laboratory instruction on the achievement in physics, science process skills and computer attitudes of 10th-grade

students. Journal of Science Education and Technology, 16(5), 451-461. https://doi.org/10.1007/s10956-007-9062-6

Yazlik, D. O., & Ardahan, H. (2012). Teaching transformation geometry with Cabri Geometry Plus II. Procedia-Social and Behavioral Sciences, 46, 5187-5191. https://doi.org/10.1016/j.sbspro.2012.06.406

Yazlık, D. Ö., & Erdoğan, A. (2019). The effect of individualized web based mathematics learning environment developed based on problem solving steps on the students' academic achievements. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 23(4), 1693-1706.

Yu, W. F., She, H. C., & Lee, Y. M. (2010). The effects of Web-based/non-Web-based problem-solving instruction and high/low achievement on students' problem-solving ability and biology achievement. Innovations in Education and Teaching International, 47(2), 187-199. https://doi.org/10.1080/14703291003718927

Zumbach, J., von Kotzebue, L., & Pirklbauer, C. (2022). Does augmented reality also augment knowledge acquisition? Augmented reality compared to reading in learning about the human digestive system?. Journal of educational computing research, 60(5), 1325-1346. https://doi.org/10.1177/07356331211062945

Additional Reports of Included Studies

Barrow, L., Markman, L., & Rouse, C. E. (2007). *Technology's edge: The educational benefits of computer-aided instruction* (Working Paper No. 2007-17). Chicago, IL: Federal Reserve Bank of Chicago.

Barrow, L., Markman, L., & Rouse, C. E. (2008). *Technology's edge: The educational benefits of computer-aided instruction* (NBER Working Paper No. 14240). Cambridge, MA: National Bureau of Economic Research.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., ... Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort* (NCEE 2007-4005). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Heppen, J. (2012). *Broadening access to Algebra I: The impact on eighth graders taking an online course* [Research brief]. Washington, DC: American Institutes for Research.

Heppen, J. B., Sorensen, N., Allensworth, E., Walters, K., Stachel, S., & Michelman, V. (2010). *Efficacy of online Algebra I for credit recovery for at-risk ninth graders: Consistency of results from two cohorts*. Paper presented at Society for Research on Educational Effectiveness conference.

Heppen, J. B., Allensworth, E., Sorensen, N., Rickles, J., Walters, K., Taylor, S. S., Michelman, V., & Clements, P. (2016). *Getting back on track: Comparing the effects of online and face-to-face credit recovery in Algebra I* (Back on Track Study Research Brief 1). Washington, DC: American Institutes for Research.

Humphry, D., & Hampden-Thompson, G. (2019). Primary school pupils' emotional experiences of synchronous audio-led online communication during online one-to-one tuition. *Computers & Education, 135*, 100–112. doi:10.1016/j.compedu.2019.03.003

Johnston, J., & Ksoll, C. (2017). *Effectiveness of interactive satellite-transmitted instruction: Experimental evidence from Ghanaian primary schools* (CEPA Working Paper No. 17-08). Stanford, CA: Stanford Center for Education Policy Analysis.

Mo, D., Zhang, L., Luo, R., Qu, Q., Huang, W., Wang, J., ... Rozelle, S. (2014). Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in Shaanxi. *Journal of Development Effectiveness, 6*(3), 300–323. doi:10.1080/19439342.2014.911770

Pol, H. J., Harskamp, E. G., & Suhre, C. J. M. (2008). The effect of the timing of instructional support in a computer-supported problem-solving program for students in secondary physics education. *Computers in Human Behavior, 24*(3), 1156–1178. doi:10.1016/j.chb.2007.04.002

Rickles, J., Heppen, J. B., Allensworth, E., Sorensen, N., & Walters, K. (2018). Online credit recovery and the path to on-time high school graduation. *Educational Researcher, 47*(8), 481–491. doi:10.3102/0013189X18788054

Robertson, D., & Miller, D. (2009). Learning gains from using games consoles in primary classrooms: A randomized controlled study. *Procedia – Social and Behavioral Sciences, 1*(1), 1641–1644. doi:10.1016/j.sbspro.2009.01.289

Sorensen, N., Heppen, J., Allensworth, E., Walters, K., Rickles, J., Taylor, S. S., & Michelman, V. (2017). Efficacy of online vs. face-to-face credit recovery courses for helping at-risk ninth-grade students in the United States get back on track. In *INTED2017 Proceedings* (p. 7375). Valencia, Spain: IATED. doi:10.21125/inted.2017.1708