Innovamat

What Impact Do Mathematics Programs Have on Learning?

How much evidence there is, what improvement can realistically be expected, and what conditions amplify it.

Carles Aulés
Research Specialist at the Innovamat Research Lab

Marc Colomer
Director of Research at the Innovamat Research Lab

Innovamat Research Series

Summary

What do we know about the impact of mathematics materials and programs on learning? Research accumulated since the beginning of this century supports a general conclusion: **curriculum materials can improve outcomes, but their effects tend to be moderate and depend greatly on how they are implemented**. Moreover, we still know far less than we would like about which programs work, for whom, and under what conditions. A review of the available evidence supports the following conclusions:

- **Rigorous evidence remains scarce.** Most of it comes from the United States, and even there, many widely used programs have never been assessed using designs capable of convincingly estimating their impact.

- **Obtaining causal evidence is costly and difficult.** A rigorous assessment may require several years, thousands of students, numerous schools, considerable funding, and schools willing to accept random assignment of programs.

- **When improvements are found, they tend to be moderate, but they can be meaningful at scale.** Meta-analyses place the average advantage of replacing one set of curriculum materials with another at approximately +0.05 to +0.10 standard deviations (SD). Among the studies reviewed by *What Works Clearinghouse* and *Evidence for ESSA*, the median is +0.13 SD. These effects are modest, but the comparison is not between receiving instruction and receiving none: the control group continues using its usual materials. These figures therefore represent the additional improvement attributable to replacing one option with another. If both have similar costs, even a moderate improvement can be important when it reaches every student.

- **The same program can produce different outcomes depending on where, how, and against what it is assessed.** The observed impact depends on the quality of the alternative used by the comparison group, the test used to measure learning, the elapsed time since the intervention, and the quality of implementation. Assessments closely aligned with the program's content tend to find larger effects than external tests, and some improvements may diminish over time.

- **Materials work better when accompanied by practice-based teacher training.** The studies synthesized by Pellegrini et al. (2021) and Lynch et al. (2019) find larger effect sizes when the intervention combines curriculum materials with ongoing support and professional development (the effect size increases by between +0.08 and 0.10 SD). The evidence suggests that providing good materials is not enough: helping teachers use them effectively increases their impact.

- **Balancing cost and benefit.** Intensive interventions such as tutoring or small-group instruction produce large effects on average. The same is true of *coaching*. However, these interventions require far more staff and are difficult to scale. Curriculum materials accompanied by training produce smaller effects per student, but they can reach virtually every classroom at a relatively low additional cost.

When we look for resources, materials, and programs to support mathematics teaching and learning, the range of options can be overwhelming. There are textbooks and other curriculum materials, supplemental digital applications, assessment tools, teacher training services, and comprehensive programs that combine materials, training, and ongoing support. These options differ greatly in their approaches and scope. Given this variety, the central question is how much difference each option can make in student learning. Does choosing one resource over another really matter? What improvement can we expect from a higher-quality curriculum?

First, it is worth clarifying the vocabulary used throughout the chapter.

- By **curriculum materials or mathematics program**, we mean a structured set of educational resources designed to support instruction in the content established by the official curriculum. These may include textbooks, teacher guides, manipulatives, digital tools, and assessment tools, among other components. To better distinguish options with different scopes, throughout the chapter we will generally use **curriculum materials** when these are offered alone or with minimal implementation support, and **program** when the materials are accompanied by more substantial implementation support, such as specific training, instructional guidance, and/or *coaching*.

- By **supplemental resources**, we mean structured materials that support instruction in specific content or skills but are not intended to cover the curriculum comprehensively: digital applications, practice workbooks, assessment tools, or other specific materials.

- By **teacher training**, we mean activities aimed at developing teachers' knowledge, skills, and overall professional capacity.

- **Intervention**: in the broadest sense, we mean any deliberate curriculum intended to improve mathematics teaching or learning. However, when we refer to a **targeted intervention**, we mean something more specific: small-group instruction that supplements classroom instruction and is generally intended for students who are struggling.

In all these cases, interpreting an impact study of an educational intervention requires a critical perspective, as we explain in Chapter 1, “How Do We Know Whether a Math Program Works?”. We must ask what type of study was conducted, what alternative it was compared with, what outcomes were measured, when they were measured, and in what context. However, assessing the rigor of a study is not enough. We also need a benchmark that allows us to interpret the magnitude of the result: even when an effect is estimated accurately, is it large or small? In practical terms, **what improvement is reasonable to expect when adopting a particular mathematics program?**

To answer these questions, we conducted an in-depth review of the available evidence. Before presenting the results, it is helpful to mark off its scope. In our review, the United States is the context in which we found the most impact assessments of mathematics programs. Other countries and regions have numerous studies on classroom practices, learning trajectories, or specific materials, but assessments of comprehensive mathematics programs are less common. This chapter therefore relies primarily on US sources. Their findings may provide a useful starting point for other education systems, but they should not be transferred automatically: their relevance will depend on factors including the official curriculum, school organization, teacher training, and the characteristics of students in each context.

Our analysis combines three sources of information. First, we examined the main meta-analyses in the scientific literature. A **meta-analysis** systematically synthesizes the results of several studies addressing the same question and makes it possible to estimate both the average effect and its variation. However, its conclusions are only as robust as the studies it includes. Second, we coded 85 studies included in *What Works Clearinghouse* and *Evidence for ESSA* as of the beginning of August 2026. These are two US repositories that review evidence on educational interventions using explicit criteria. Third, we collected 229 documents that six major US publishers presented on their websites as evidence for their products. These documents are not necessarily equivalent to independent impact studies; for precisely this reason, we analyzed the type and strength of evidence each provided.

This review yields **six conclusions**, which we explore throughout the chapter. Rigorous impact evidence is the exception, even in the United States. Generating it is slow, expensive, and difficult. When such evidence exists, the effect tends to be modest, although it may represent an efficient investment. The same curriculum materials do not always have the same impact because outcomes depend on which test is used and when and under what conditions they are assessed. The impact depends not only on the quality of the materials but also on what accompanies them: training, ongoing teacher support, and the conditions needed to use them effectively in the classroom. Compared with other levers such as tutoring or *coaching*, curriculum materials accompanied by training have a smaller effect per student but a much broader reach: there is a trade-off between impact and scale.

By the End of This Chapter,

you will know what proportion of mathematics programs have strong impact evidence and why generating it is so difficult; what improvement can reasonably be expected from changing mathematics programs—generally modest, though potentially meaningful; why the same curriculum materials do not always produce the same result; what conditions can amplify their effect, particularly teacher training and ongoing support; and what trade-off exists between the impact and scale of the different levers a school can use.

Effect size. The difference between the average outcome of the group that received a program and that of the comparison group, expressed in standard deviations—SD for short. In a well-designed causal assessment, this difference can be used to estimate how much the outcome changed because of the program.

In the article "How Do We Know Whether a Math Program Works?" we explain in greater detail how to interpret an effect size qualitatively, but here it is enough to remember two ideas.

The first is that we need a realistic benchmark, not a universal rule. In Kraft's expanded review, the median of 3,426 estimates from 973 randomized trials was 0.10 SD. As a starting point for causal studies of educational interventions assessed using standardized achievement tests, Kraft suggests considering effects below 0.05 SD small, effects from 0.05 to less than 0.20 SD medium, and effects of 0.20 SD or more large. These categories provide guidance, but they do not replace an analysis of the design, measurement used, population, cost, or feasibility of implementing the program at scale (Kraft, 2020, 2023).

The second is that the same SD value does not represent the same proportion of typical annual progress in every grade. In the historical US standardized test data analyzed by Hill et al., annual gains measured in SD are larger in the early grades and smaller in middle and high school. If, as an approximation, we divide this annual progress by nine months, 0.10 SD is equivalent to about one month of observed progress in the early elementary grades and approximately four months in grades 8 through 10. This equivalence helps build intuition, but it should not be interpreted as a literal measurement of months of learning: it depends on the grade, the test, and several assumptions about how progress is distributed throughout the year (Hill et al., 2008; Kraft, 2020).

1 | Rigorous Impact Evidence Is the Exception, Not the Norm

How Common Is Rigorous Impact Evidence?

In 2023, Bellwether examined the available evidence for 59 mathematics programs from US publishers. Before continuing, try to answer this question: for how many of these programs do you think they found at least one efficacy study that met minimum standards of scientific quality?

The answer is 17. For the other 42—approximately seven out of ten—Bellwether found no studies that met all its requirements. Among other criteria, a study had to have been conducted in the United States, focus on K–9 students, include at least 250 participants, and be published in one of the formats accepted by the review. It is worth noting that these 59 programs were not selected randomly. They were 59 programs that EdReports—an independent nonprofit organization that assesses instructional materials and publishes expert reviews—identified as being of medium or high quality. In other words, these programs have a seal of quality for their curriculum content. If we expanded the list to include other materials and programs, the percentage of programs with available impact studies would probably be even lower.

But the story does not end there. The US Every Student Succeeds Act (ESSA) distinguishes four **levels of evidence**. Level 1, known as strong evidence, requires at least one favorable and statistically significant result from a well-designed and well-implemented experimental study. In this type of study, called a **randomized controlled trial** (RCT), assignment to the program or comparison group is random. Level 2, or moderate evidence, is based on rigorous quasi-experimental studies; Level 3, or promising evidence, is based on correlational studies that incorporate statistical controls to reduce selection bias; and Level 4 requires at least a theoretical rationale accompanied by some effort to assess its effects. The question is, among the available studies, how many conducted a randomized trial aligned with ESSA Level 1?

Bellwether found 101 studies of these 17 programs and retained 61 after applying its criteria. Of these, approximately one in ten used a randomized design. Most of the remaining studies used **quasi-experimental**, correlational, or other approaches. In global terms, the United States is the country with the most research on the impact of mathematics programs. Within the United States, most of these programs have no studies assessing their impact. Among those that do, the vast majority use designs that do not support the same degree of confidence as an experimental study that the observed impact was caused by the intervention.

To supplement these findings, we conducted the reverse exercise: instead of starting with studies included in evidence catalogs, we examined what a school finds when it visits **providers' own research pages**. As of July 2026, we collected the documents that six providers with a significant presence in the US market—HMH, Curriculum Associates, McGraw Hill, Savvas, Great Minds, and Amplify—presented on their websites as research or evidence. In total, we coded 229 documents associated with 31 mathematics programs (Innovamat Research Lab, 2026b).

Only 7% of these documents were randomized trials. Approximately six out of ten did not include an external comparison group. For example, they presented outcomes before and after using the program, descriptions of its implementation, or associations between the degree of use and outcomes. These correlational analyses provide useful information. Because organizing a randomized trial is extremely complex and expensive, most published work can be expected to consist of simpler analyses, which also provide useful information about how a program is implemented and how its use relates to outcomes.

That said, it is important to understand what each type of study can and cannot answer. A **correlational analysis** is simpler but also less rigorous: on its own, it cannot estimate what would have happened with the same students if they had not used the program, and without a credible **counterfactual**, observing an improvement is not enough to attribute it to the intervention. In addition, as of July 2026, 23 of the 31 programs had no impact rating from *What Works Clearinghouse* or *Evidence for ESSA*. In other words, no external organization had reviewed their studies to classify them by level of evidence. None of this is unique to these six publishers. It describes the state of the sector as a whole. Generating causal evidence for a program is slow, expensive, and difficult.

"Generating causal evidence for a program is slow, expensive, and difficult."

Of 59 US mathematics programs highly rated by EdReports, 42 did not have a single efficacy study published in 2023 (Bellwether, 2023). On the websites of the six major publishers, most studies are correlational. In other words, these studies cannot detect a program's causal effect on students' academic outcomes.

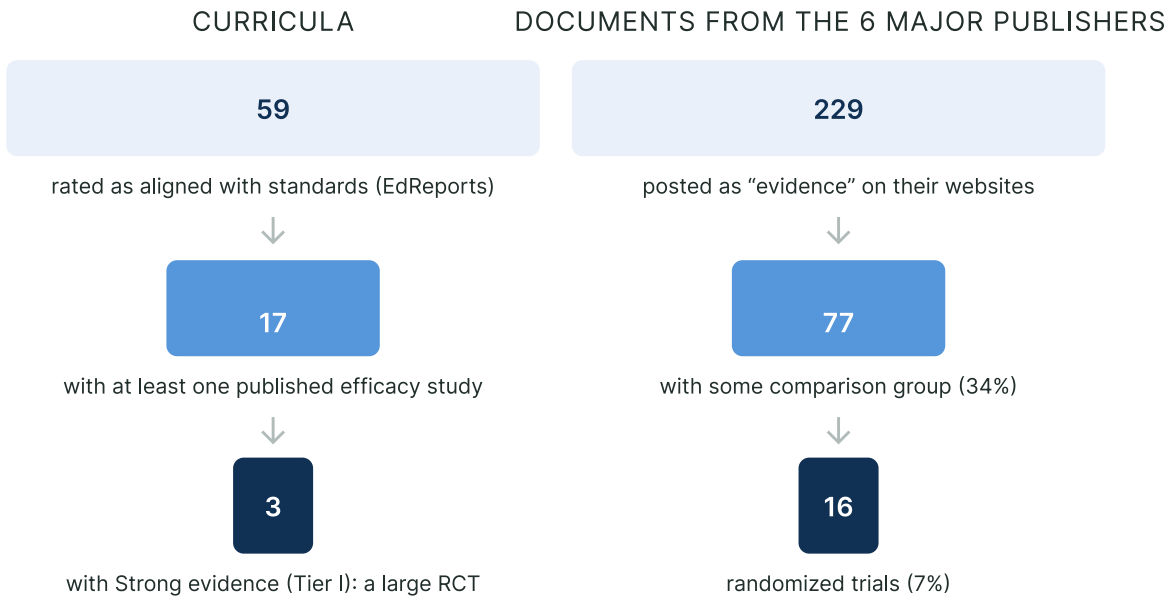


Figure 1. On the left, the curriculum programs assessed by EdReports, according to the Bellwether study (2023). On the right, what the six major publishers present as evidence, according to the Innovamat Research Lab study (2026a, 2026b).

Why Are There So Few Studies With an Experimental Design?

There is no single cause. The most common explanations include the following:

- **Cost.** A large-scale independent assessment can cost millions: the *Cognitive Tutor Algebra I* assessment received nearly $6 million from the Institute of Education Sciences (2007). In most cases, experimental studies are publicly funded.
- **Time.** A serious trial requires several school years for design, implementation, and analysis. The same *Cognitive Tutor Algebra I* study found no effects in the first year but did find effects in the second (Pane et al., 2014).
- **Access to schools.** Researchers must persuade dozens of schools and districts to let chance determine which schools or classrooms use the program. This means that not all participants can begin using it from the outset, which may create resistance and logistical complications, such as managing different instructional approaches simultaneously. It may also be difficult to secure family cooperation and prevent teachers from sharing materials or practices between the intervention and control groups. This could reduce the differences between the groups and make it more difficult to assess the program's true impact.
- **Coordination and independence.** To increase independence, an external party should design, conduct, and publish the assessment.
- **Incentives.** Sellers have little incentive to subject an existing commercial product to a test with an uncertain outcome (Bellwether, 2023).

Strong empirical evidence is the exception rather than the norm, even in the United States, the country with the most research on the impact of mathematics programs. Most US mathematics programs have no studies assessing their impact. Among those that do, the vast majority use designs with limitations that prevent us from concluding with complete confidence that the observed impact is caused by the program assessed. Correlation is not causation.

2 | When Evidence Exists, the Effect Is Usually Modest

When a math program is assessed using a design capable of estimating its impact, what difference does it usually make? When the assessment focuses on specific curricular materials rather than a comprehensive program that also includes professional development and support, the main reviews point to a small but meaningful improvement.

Slavin and Lake (2008) analyzed 13 studies of curricular approaches—mostly textbooks—and estimated a mean effect of +0.10 SD. Years later, Pellegrini et al. (2021) applied stricter inclusion criteria and obtained a smaller, nonsignificant estimate: +0.05 SD across 30 studies of traditional and digital curricular materials that included little teacher professional development. Although the two reviews differ in the period studied and their methodological criteria, they provide estimates of a similar magnitude. Therefore, the average advantage of replacing one set of curricular materials with another appears to be around +0.05 SD and, in some syntheses, close to +0.10 SD. These figures compare the curricular materials assessed with those the school was already using, which also teach mathematics, and do not isolate the quality of the materials: studies of real-world textbook adoptions find few average differences among them, even when they have different quality ratings (Blazar et al., 2020).

Large-scale studies of real-world textbook adoptions provide an equally nuanced picture. After combining data from six US states following the implementation of the Common Core State Standards, Blazar et al. (2020) found few average differences among the materials analyzed. In contrast, Koedel et al. (2017) estimated that one of the four most widely used textbooks in California outperformed the others by approximately +0.05 to +0.08 SD in grade 3, with favorable results in later grades as well. Overall, these studies suggest that differences among curricular materials may exist, but they do not necessarily appear in every context or always reach the same magnitude.

The Available Evidence Is Not a Neutral Sample

The figures above describe the mean effect found in certain syntheses of scientific literature, but the evidence available to us is far from a neutral sample of all the studies conducted. Some results are more likely than others to be published, included in an evidence clearinghouse, or highlighted by their authors.

This pattern is clear in our analysis. We coded the 85 studies of the 52 math programs in the *Evidence for ESSA* and *What Works Clearinghouse* catalogs. They include all types of programs, not only math programs: supplemental resources (apps and digital platforms), targeted interventions and tutoring, teacher professional development, and school-wide reforms; 11 of the 52 are math programs. None of the 85 studies reported a negative main effect. Among the 66 that report a usable effect size, the median is +0.15 SD. By program type, the median is +0.13 SD for both the 15 studies of math programs and the 27 studies of apps and platforms, +0.28 SD for the 17 studies of targeted interventions and tutoring, and +0.08 SD for the 7 studies of teacher professional development and school-wide reform (Innovamat Research Lab, 2026a). These figures are somewhat higher than those obtained in several of the previous syntheses. Even more striking is the absence of any negative main result.

Part of this difference is explained by how the clearinghouses themselves are structured. ESSA tiers 1, 2, and 3 require at least one favorable, statistically significant result and no relevant negative results. Therefore, a sample made up of programs that reach these tiers selects interventions that have produced at least one positive result from the outset. It would be incorrect to interpret their mean effect as representative of all available math programs.

The contrast with the study by Lortie-Forgues and Inglis (2019) helps illustrate this selection. The authors analyzed 141 large, publicly funded randomized trials in the United States and the United Kingdom involving more than 1.2 million students. The interventions assessed were highly diverse and had not been selected because they had previously produced favorable results. Their mean effect was +0.06 SD. In addition, 40% of the trials were inconclusive: the data did not clearly tip the scales either toward the existence of an effect or toward its absence.

Publication bias compounds the selection inherent in the clearinghouses. It occurs when studies with positive or statistically significant results are more likely to be published than those that find null or negative effects. In an analysis of 645 studies from 12 reviews of educational research, Cheung and Slavin (2016) found that the effects reported in published articles were, on average, approximately twice those reported in unpublished documents.

The visible evidence may therefore be shaped by several related but distinct filters: which studies are published, which are included in clearinghouses, and which results are selected as the main findings. This does not invalidate the observed effects, but it requires them to be interpreted with caution. The +0.15 SD in our database, or the +0.13 SD for math programs, describes the studies selected for these clearinghouses, not necessarily the effect that can be expected from any math program.

The Same Intervention Does Not Always Have the Same Impact

As we saw in Chapter 1, **effect size** is not a fixed label that can be assigned to a program regardless of the circumstances. In addition to depending on the alternative with which it is compared, it varies according to what is measured, when it is measured, and the conditions under which it is implemented. In other words, there is no single abstract "program effect," but rather effects estimated in specific contexts and at specific times.

The first factor to consider is **the assessment used**. When a program describes its **theory of change**, it is essential to consider what we expect it to change or improve and what we do not. In this regard, the assessment used to evaluate the program is crucial. When an assessment measures content, tasks, or representations that closely resemble those addressed during the intervention, the effects are usually larger than when more general assessments are used, which require students to apply what they have learned to less familiar situations (Cheung and Slavin, 2016). Using more general assessments provides a better measure of the extent to which learning transfers beyond the activities practiced.

The *Fusion* study provides a good example. The effect was g = 0.77 on *ProFusion*, an assessment developed by the researchers to measure content closely related to the intervention. On more general assessments, the effects were smaller, ranging from 0.02 to 0.20 (Clarke et al., 2023). This difference does not mean that students learned nothing, since it may reflect both limited transfer at that point and a general assessment that was not sensitive to what the intervention addressed. Results on the aligned assessment indicate whether students learned what the intervention was intended to teach, while independent assessments indicate whether that learning transfers to other types of tasks.

Related to this point, our analysis found a lack of measures that go beyond learning (Innovamat Research Lab, 2026b), such as attitudes toward mathematics, mathematical identity, or staying in school. Published studies focus on learning measured through standardized assessments that, in many cases, capture only part of what a program aims to change.

"Published studies focus on learning measured through standardized assessments that, in many cases, capture only part of what a program aims to change."

A second factor to consider is when the results are assessed. Some interventions produce a positive impact immediately after they end that decreases over the following months or years, a phenomenon known as ***fade-out***. Research indicates that this loss of impact is common in certain interventions targeting specific cognitive skills, especially when they are brief and are no longer implemented (Bailey et al., 2017, 2020). However, there is no universal proportion of the effect that is retained: its development depends on the type of intervention, the learning assessed, and students' subsequent experiences.

The one-on-one tutoring program *Math Recovery* illustrates this phenomenon. At the end of grade 1, it produced effects ranging from +0.15 to +0.30 SD on assessments administered by external evaluators and even larger effects on an assessment closely aligned with the program. By the end of grade 2, however, the estimates were small, and none were statistically distinguishable from zero (Smith et al., 2013). This does not prove that no learning remained. Rather, one grade later, the study could no longer establish a reliable advantage for the group that received the intervention over the comparison group.

The third factor is the **implementation conditions**. We can think of curricular materials as a musical score and teachers as the performers. Without a performer, there is no music, and the same performer does not get the same results from every score. Some materials may support richer tasks, more coherent learning progressions, better connections among ideas, and more opportunities for practice and feedback. Others may make all of this more difficult. However, the result also depends on how teachers understand the curriculum, how much time they can devote to it, how they respond to students' needs, and what professional development and support they receive to implement it in the classroom.

Therefore, the effect observed in a study is not solely the result of the materials. It is the combined result of the program, the people using it, and the organizational conditions that make its implementation possible. Two schools may adopt the same curricular materials and obtain different results if they differ in the time available, teacher support, staff stability, or the extent to which they incorporate the curriculum into everyday practice. We discuss these aspects in greater depth in the article "What Does It Take to Learn Mathematics?".

"Two schools may adopt the same curricular materials and obtain different results if they differ in the time available, teacher support, staff stability, or the extent to which they incorporate the curriculum into everyday practice."

In summary, the impact of the same intervention depends on four things: what it is compared with (the more similar the comparison group and its activities are, the more difficult it will be to detect a positive effect), the assessment used (a program-specific assessment aligned with the program will produce a larger effect than an external, general assessment), when it is assessed (the more time that passes after the intervention, the smaller the measured effect usually is), and how it was implemented.

This issue becomes especially important when an intervention is implemented at scale. A program may obtain good results under especially favorable conditions—with selected schools, close monitoring, and intensive support from its developers—but fail to reproduce results of the same magnitude when expanded to more diverse contexts and implemented with the system's usual resources. For this reason, guidelines from the Institute of Education Sciences and the National Science Foundation distinguish between **efficacy** and **effectiveness** studies. Efficacy studies analyze whether an intervention can work under favorable and relatively controlled conditions. Effectiveness studies examine whether it continues to produce results under ordinary conditions, with diverse populations and no more involvement from the developers than would be expected in typical implementation (IES and NSF, 2013).

Final Notes

The recurring idea in this section is that the same intervention does not always have the same impact: the effect measure should not be interpreted as a fixed property of a program, but in relation to the question the study aims to answer: What difference does using this program instead of the available alternative make for a specific outcome, population, and context? Therefore, it is important to remember something shared by many studies of curricular materials. The comparison group does not stop receiving instruction but continues to use its usual materials. Thus, the effect size does not represent everything students learn with the new program, but rather its additional advantage over another curriculum through which they would also be learning mathematics. This rigorous comparison helps explain why effect sizes in these studies range from +0.05 SD to +0.10 SD. According to the benchmarks proposed by Kraft (2020, 2023), +0.10 SD falls in the middle range of effects observed in causal studies of educational interventions measured using standardized assessments.

This leads to the key question: Is this improvement enough? Is it worth the investment? As Whitehurst (2009) argues after comparing the costs and benefits of multiple educational interventions, changing curricular materials can be an efficient lever because the most effective programs do not necessarily cost more to acquire than the alternatives. Modest does not mean irrelevant. Choosing one set of curricular materials over another barely changes the cost, but it may produce a moderate improvement that reaches every student.

“Modest does not mean irrelevant. Choosing one set of curricular materials over another barely changes the cost, but it may produce a moderate improvement that reaches every student.”

At best, the typical effect of changing curricular materials ranges from +0.05 to +0.10 SD, although this effect may decrease over time or when the intervention is implemented at scale. Even so, changing curricular materials can be an effective lever because alternative programs often have very similar costs.

3 | Curricular Materials Deliver Results When They Do Not Travel Alone

Curricular Materials Deliver Better Results When Paired With Professional Development

Changing the materials used in the math classroom produces only a small effect on its own. As we have already seen, the meta-analysis by Pellegrini et al. (2021) places the impact of curricular materials provided with little professional development at +0.05 SD, a nonsignificant effect. When those same materials are accompanied by professional development and implementation support, the effect rises to +0.12 SD. The meta-analysis by Lynch et al. (2019), covering 95 studies of mathematics and science, points in the same direction: programs that provide only materials or only professional development yield around +0.15 SD, while those combining both reach +0.25 SD. Materials alone do little to change learning. Improvement occurs when they are accompanied by professional development specific to those materials.

The two figures are not interchangeable. Pellegrini's figure (+0.12 SD) comes from seven Elementary studies focused only on mathematics in which professional development accompanied a traditional textbook, and it is compared with +0.04 SD for those same materials with little professional development. Lynch's figure (+0.25 SD) combines mathematics and science, all grade levels, and programs that pair professional development with materials of any kind. It also includes assessments designed by the researchers themselves, which yield effects around 0.27 SD larger than standardized assessments. What both analyses share is the difference: **professional development tied to curricular materials** adds between +0.08 and +0.10 SD compared with materials alone.

It is worth clarifying what these meta-analyses mean by professional development. In Pellegrini et al. (2021), curricular materials "with little professional development" are those for which teachers receive less than two days or 15 hours of training, amounting to little more than an introduction to the materials. The category that reaches +0.12 SD includes programs with at least two days or 15 hours of professional development devoted to implementing specific curricular materials in the classroom, followed by *coaching* during the school year. In Lynch et al. (2019), professional development includes summer workshops, follow-up sessions during the school year, *coaching*, and collaboration among teachers at the same school. The curricular materials are programs that introduce new tasks and sequences with teacher guidance embedded in the materials. One example of a program combining both is a fractions unit accompanied by a five-day summer workshop and 13 school-based sessions during the school year (Saxe et al., 2001). Another is *SimCalc*, a Middle School proportionality unit with software and a professional development workshop for teachers that was evaluated in three large-scale trials in Texas (Roschelle et al., 2010). In Lynch's analysis, professional development delivers better results when it focuses on how to use the materials and how students learn the content, includes implementation follow-up meetings, and involves teachers from the same school. In contrast, the number of hours does not predict the effect, and online components are associated with smaller effects.

One might think that teacher professional development is what truly creates the impact. However, three large randomized trials partly contradict this intuition: intensive professional development in mathematical content that was disconnected from the classroom increased teachers' mathematical knowledge without producing any detectable positive effect on their students (Garet et al., 2016). In the two earlier trials in the same series, it did not even improve teacher knowledge (Garet et al., 2010, 2011). The effect of professional development appears to reach students when actual practices in the math classroom change (Lynch et al., 2025). Professional development that encourages teachers to think about their own classrooms and provides resources they can put into practice has a greater impact than professional development limited to prescribing behaviors or adding content without real examples teachers can adapt to their classrooms (Kennedy, 2016).

"The effect of professional development appears to reach students when actual practices in the math classroom change."

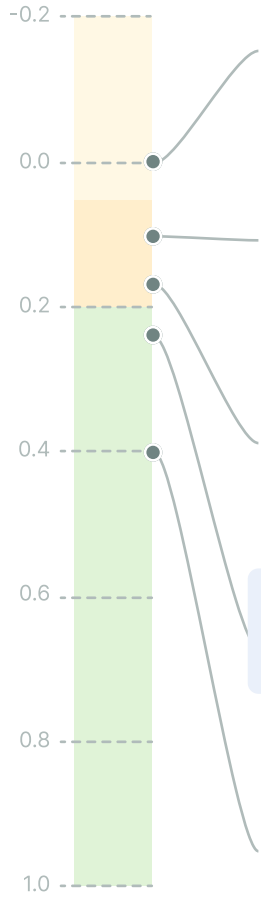
How It Compares With Other Levers: Effect, Cost, and Scale

So far, we have compared the same curricular materials with and without professional development. The second question is different: How does that combination compare with other ways a school can pay to improve learning, such as *coaching* or tutoring, in terms of effect, cost, and scale?

The first alternative is ***coaching***: one-on-one classroom support from an expert who observes and plans lessons with the teacher over several months (as we explain in greater detail in Chapter 3). It is the form of teacher professional development with the largest known effect on classroom practice: +0.49 SD (Kraft et al., 2018). Its effect on student learning is smaller, +0.18 SD, and comes almost entirely from reading interventions; in mathematics, the evidence is still limited and nonsignificant. In addition, it should not be viewed as a lever separate from the materials: the professional development tied to curricular materials that reaches +0.12 SD in Elementary includes *coaching* as a component. Returning to our musical score analogy, general professional development is a short course in conducting an orchestra; *coaching* is rehearsing week after week with a master beside the music stand.

Building Instructional Capacity Optimizes for Effectiveness and Scalability

Better materials and better teaching support compound. The more the system is built around empowering the teacher, the further the gains reach.



Types of intervention	Average Effect Size	Scalability	Assessment
PD alone (no link to materials or classroom)	**~ 0.00 SD** Garet et al. (2010, 2011, 2016); Pellegrini et al. (2021)	High	Good training changes what teachers know, not how they teach.
HQIM (core materials, minimal training)	**+0.05 to +0.10 SD** Pellegrini (2021); Slavin & Lake (2008)	High	Good materials change what teachers have, not how they use them.
Coaching (one-to-one instructional coaching)	**+0.10 to +0.18 SD** Kraft et al. (2018)	Low	Good coaching changes teacher effectivity, but is labor intensive.
HQIM + HQPL (materials + curriculum-linked PL)	**+0.12 to +0.25 SD** Pellegrini (2021); Lynch (2019)	Medium	Good materials taught the right way, change what a teacher does.
Tutoring / Intervention (small-group, high-dosage support)	**+0.29 to +0.40 SD** Nickow (2024); Kraft (2026)	Low	Good tutoring is highly effective, but hard and expensive to scale.

Interpreting Effect Sizes (Kraft, 2020) • Small = < 0.05 SD • Medium = 0.05 to < 0.20 SD • Large = ≥ 0.20 SD

* For visualization purposes, each intervention is plotted at the highest effect size observed in the reviewed literature; lower estimates from the same studies appear in the Average Effect Size column.

Figure 2. Comparison of the effect size and scalability of different educational interventions, ordered from least to greatest estimated effect. **Note**: digital practice platforms are excluded from the graph due to lack of comparability: most effect sizes reported consider students that use them with fidelity, not all students in the classroom. Regarding coaching, most programmes in Kraft et al. (2018) are reading coaching, not math.

The second alternative, and the one with the greatest direct effect on students, is **tutoring** in very small groups during school hours. A meta-analysis of nearly 100 randomized trials found a mean effect of +0.29 SD across tutoring programs (Nickow et al., 2024), while the broader synthesis of 263 trials found an effect of +0.40 SD (Kraft et al., 2026). It is the most powerful intervention known to educational research.
Tutoring and *coaching* have something in common. They are very expensive because they require highly individualized intervention, extensive human resources, and sustained involvement over time. This makes the solution difficult to scale across many contexts. The more an intervention focuses on an individual student, the more resources it requires and the more difficult it is to scale (Figure 2). Tutoring requires one person for every three or four students. *Coaching* requires one expert for every few teachers, and its effects decrease by half or more as programs grow: from +0.28 SD in pilot programs to +0.10 SD at scale (Kraft et al., 2018). When tutoring is measured in large-scale implementations using external assessments, its effect decreases to +0.16–0.22 SD (Kraft et al., 2026). Curriculum materials follow the opposite path: their effect in isolation is the smallest, but they reach every classroom at once, every day, at almost no cost. Training connected to the materials, which teaches educators to use the specific curriculum materials available in their classrooms, strikes a balance between effect and scale: it increases the effect of the materials without sacrificing their scalability. Only tutoring clearly outperforms this combination; *coaching* matches it, but at a much higher cost per teacher.

> "The more an intervention focuses on an individual student, the more [resources] it requires and the more difficult it is to scale."

The Instructional Core

Let us pause for a moment to consider what all these findings have in common. Effective training changes what happens in the classroom. Effective tutoring is connected to the curriculum materials used every day. Effective materials are used by someone who knows how to use them. Everything points to the same conclusion: learning does not occur in the materials, the teacher, or the student separately, but through the interaction among all three. This is what the literature defines as the **"instructional core"** (Figure 3).

Learning occurs within the "instructional core": the interaction between teachers and students around the materials. Given this interaction, improving one factor while ignoring the other two accomplishes little. The impact is greater when the intervention or program considers all three.

Figure 3. The instructional core is an interaction among three factors: if one factor fails, the set accomplishes little. Based on Cohen et al. (2003).

The same logic applies outside the United States. The World Bank expert panel identifies **"structured pedagogy"**—quality materials combined with teacher training and support—as one of the most cost-effective educational investments available (GEEAP, 2023). It is not a specific program; it is precisely the combination discussed in this section.

Let us summarize what we have learned in this section. First, changing curriculum materials can produce small or moderate effects. Second, when these changes are accompanied by practice-based training, effect sizes increase. Third, there is a trade-off between **impact and scale**: tutoring and small-group interventions have the greatest effects, while *coaching* provides the most intensive training, but these approaches are also the least scalable. Their role is therefore to complement, not replace.

But this is where the orchestra analogy reaches its limit: in a school, the "audience" also performs. Students do not simply receive the program; they participate in it from their own starting points and within their family and social contexts. There are also more musicians surrounding the classroom than it may seem: the school leadership that determines schedules and priorities, the education policies that establish the official curriculum and resources, peers, and families. This chapter has examined the score and its rehearsals. Chapter 3 broadens the view to the whole picture: which factors determine mathematics learning, which can be changed and which are predetermined, and what each stakeholder can do to make the performance sound as good as possible (Chapter 3, "What Does It Take to Learn Mathematics?").

Curriculum materials produce effects of between +0.05 and +0.10 SD; with practice-based teacher training, they produce effects of between +0.12 and +0.25 SD, depending on the synthesis. Tutoring and small-group interventions have greater effects, and *coaching* primarily improves teaching practice, but both cost more and are more difficult to scale: they serve as complements, not standalone solutions. Programs that combine good materials with practice-based training strike the best balance between effect and scale.

4 | What This Means in Practice

1. Adjust expectations. A sustained effect of 0.10–0.20 SD is among the best this field has to offer: students reach the same point several weeks or months earlier. These are meaningful but gradual improvements, not radical transformations within a single school year. Educational gains tend to be modest, cumulative, and context-dependent, so promises of dramatic change should be evaluated with caution.

2. Request studies and ask questions. Impact studies and evidence help us make better-informed decisions. It is important to request evidence from developers and analyze it carefully. What is the effect size, and how does it compare with the benchmark of 0.10 SD? If studies do not exist or provide little transparency about their design and effect measurement, ask why. It is also important to think critically when reading studies, as we saw in Chapter 1: Who authored the study—the developer or an external evaluator? What design was used, and how did the researchers estimate what would have happened without the program? Which assessment was used to evaluate the program?

3. Evaluate the complete program package, not just the materials. The training and support provided with the materials may matter more than the materials themselves. In the meta-analysis by Lynch et al. (2019), programs that include only materials or only training produce effects of about +0.15 SD, while those combining both reach +0.25 SD. The elementary studies by Pellegrini et al. (2021) show a similar contrast: +0.05 SD with little training and +0.12 SD with implementation support. The two syntheses measure different things—elementary education and external assessments in one case, and mathematics and science using researcher-developed assessments in the other—but agree on the key point: training connected to curriculum materials adds between +0.08 and +0.10 SD. Asking what type of training is included should be as standard as reviewing the teacher guides or student books.

5 | Frequently Asked Questions

"Few programs have strong evidence. Which ones do?"

Three mathematics programs currently hold the highest rating—strong evidence—in the *Evidence for ESSA* catalog. *Math Expressions*: +0.11 SD in the largest publicly funded curriculum trial in the United States, with more than 100 schools randomly assigned and outcomes measured using an external assessment (Agodini et al., 2010, 2013). *Math in Focus*: approximately +0.12 SD in a randomized trial in Nevada commissioned by the publisher from an external research firm, with no effect on the state assessment (Jaciw et al., 2016). *Carnegie Learning*'s algebra program: no effect in the first year and approximately +0.20 SD in the second year in RAND's seven-state trial (Pane et al., 2014).

"The program we are considering does not appear in the What Works Clearinghouse. Is that a bad sign?"

Not necessarily. To begin with, programs from outside the United States are rarely listed, simply because certification requires studies conducted within the United States. In this case, check whether the program has any studies published in a scientific journal and analyze how the comparison group was defined, which assessment was used, when the study was conducted, and in what context. Furthermore, the two major US registries—*What Works Clearinghouse* and *Evidence for ESSA*—overlap far less than it may seem: of the 52 programs in the *Evidence for ESSA* catalog, only 3 also have a *What Works Clearinghouse* report. Conversely, some programs with good evidence are absent from the catalog. *Building Blocks*, an Early Childhood Education program, has peer-reviewed randomized trials but has not been certificated. There are also telling omissions: the WWC reviewed *enVision*, one of the country's best-selling programs, and did not find a single study that met its standards.

"Have you analyzed the evidence from major US publishers? What did you find?"

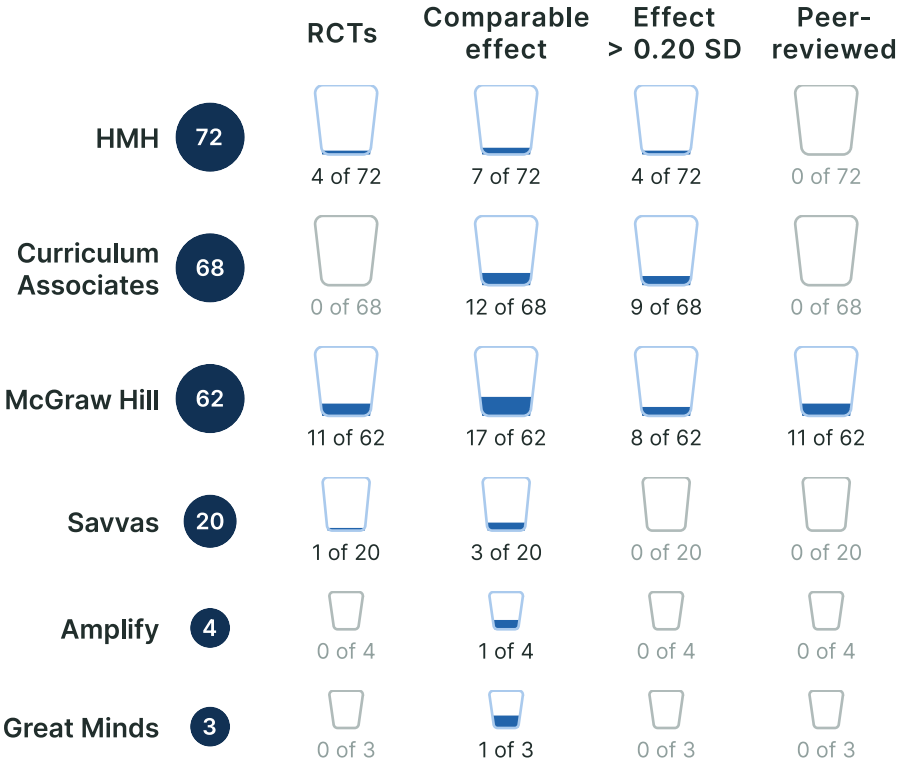
Yes, we conducted an in-depth analysis of the publicly available evidence from six major publishers. We collected and individually coded the 229 documents that six major publishers presented as research or evidence on their own websites as of July 2026. We describe the type of evidence each publisher chooses to post on its website, not the quality of its programs: a publisher may have a good program but communicate little about its impact, while publishing a large volume of material does not guarantee that the material can be used to estimate an effect.

HMH publishes the most, with 72 documents on its research website. More than half are before-and-after comparisons of program use without a comparison group. Curriculum Associates publishes 68 documents, about 60% of which are correlational analyses, and none are randomized trials. McGraw Hill publishes 62 documents and is the only publisher in the group with peer-reviewed articles, concentrated in two programs: the randomized trials of *Building Blocks* found substantial positive effects and are among the strongest curriculum studies published in academic journals, while the three *ALEKS* studies found no statistically significant effects. Savvas publishes 20 documents, almost all with comparison groups, although only 3 include an effect size—the measurement that allows us to assess the magnitude of an impact and compare it with other options. Great Minds publishes 3 documents about *Eureka Math*, while Amplify publishes 4. A smaller volume does not necessarily indicate a weaker evidence base, as it may reflect a more selective publication policy. Overall, 7% of the documents are randomized trials, 18% report a standardized effect size relative to a comparison group, and 5% have undergone peer review. The complete database, including the results of each study, is openly available (Innovamat Research Lab, 2026b) (Figure 4).

These figures should be placed in context before drawing conclusions. A randomized trial under real-world conditions is costly and time-consuming: it requires persuading dozens of schools to let chance determine which program they will use, sustaining the study over several school years, funding an independent evaluation, and accepting in advance that the result may not be favorable. RAND's evaluation of *Cognitive Tutor Algebra I* across seven states cost nearly $6 million and lasted for years (Institute of Education Sciences, 2007; Pane et al., 2014). No publisher can launch a study like this on its own, and even if it wanted to, the cost would often be prohibitive. Furthermore, our count describes what each publisher had posted on its website on a specific date, not everything that exists: there may be ongoing studies, unpublished research, or third-party evaluations that do not appear on these websites. Once a study has been completed, publishing it in peer-reviewed scientific journals makes the process even slower and more complex. In fact, many valid impact studies are never published in an academic journal.

The Major Publishers: Many Studies, Very Little Comparable Evidence

229 documents located (August 2026); only 41 measure impact with a standardized effect size relative to a comparison group, and only 11 have undergone peer review. Each cube fills according to the proportion.



"Comparable effect" = standardized effect size (d or g) from a design with a comparison group; pre/post designs without a control group, beta coefficients, and percentage points are excluded. At Curriculum Associates, most effects > 0.20 SD are measured using the publisher's own assessment (i-Ready Diagnostic).

Source: Innovamat Research Lab (2026b); 229 documents from the research websites of the six major US publishers, collected in August 2026.

Figure 4. Documents about 31 mathematics programs published on the research websites of six major publishers (229 documents, July 2026). Publishers post many documents but very little comparable evidence: only 41 of the 229 report a standardized effect size relative to a comparison group, and only 11 have undergone peer review. At Curriculum Associates, most large effects are measured using the publisher's own assessment. Authors' analysis (Innovamat Research Lab, 2026b).

"Isn't a study by the publisher better than no study at all?"

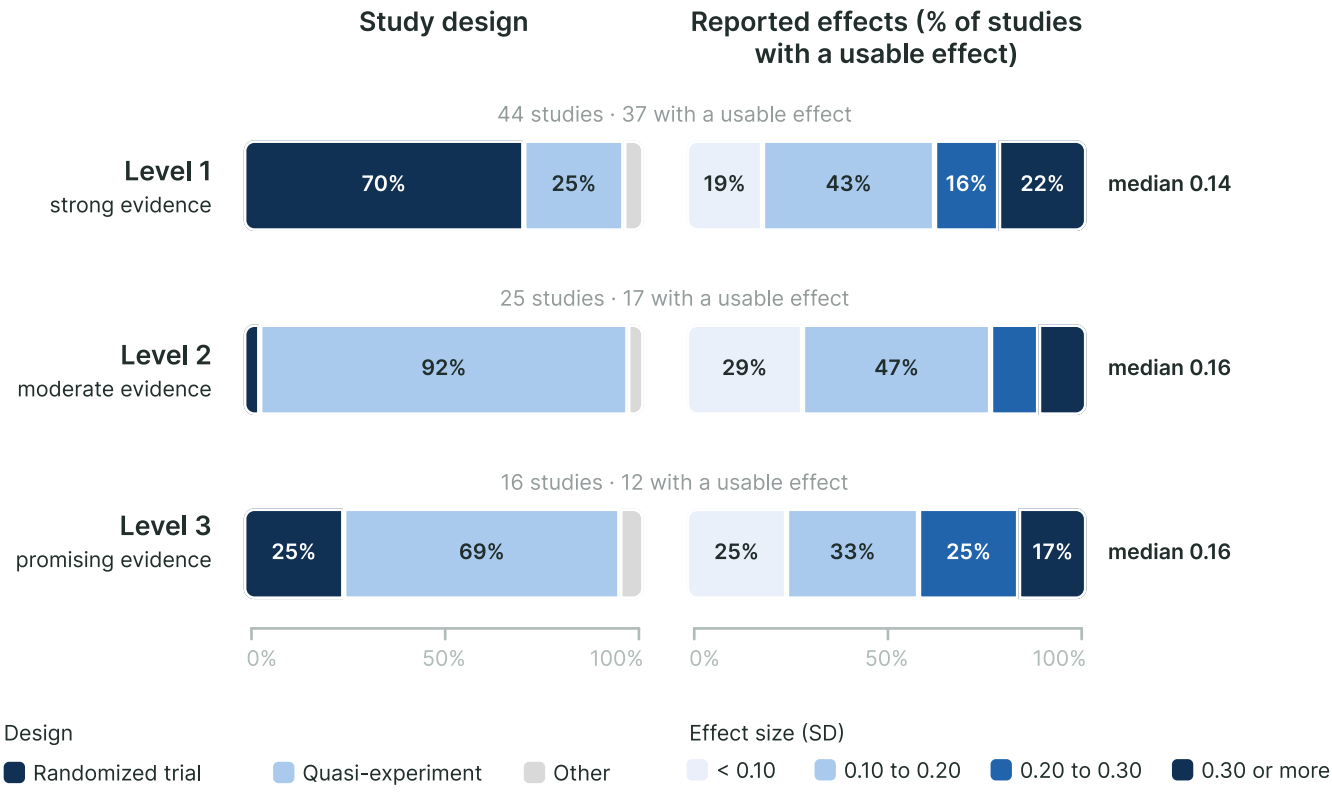
Yes, as long as it is read with an understanding of what it is. Data from some research studies reveal two things about internal studies. First, they report less information and do so less effectively—almost half provide no effect size, only claims of improvement. Second, they tend to use their own assessment, aligned with the content taught, which consistently produces larger figures than an external assessment. The studies themselves show this clearly because they administer both assessments to the same students over the same period: *Math 180* reports 0.51 SD on the publisher's assessment and 0.32 on the external assessment; *Math Recovery* reports 1.04 on its own measure and 0.28 externally; *ROOTS*, 0.81 compared with 0.23; and *Fusion*, 0.77 compared with 0.07, with the latter not statistically significant. However, a difference like this does not prove that no learning occurred: the learning may not yet transfer to less familiar tasks, or the much broader general assessment may not be sensitive enough to detect it. This is why it is essential to know which assessment is being considered.

"How are study designs and effects distributed within the certified catalog?"

There is substantial variation, particularly among evidence levels. Our database covers 52 curriculum programs of all types and 85 studies from the *Evidence for ESSA* mathematics catalog. Of these studies, 44% are randomized trials, 53% are quasi-experimental studies, and 3% use other designs. The certification level predicts the type of design almost perfectly: among studies of Level 1 programs, 70% are randomized—not 100%, because the rating applies to the program, not the study; one strong RCT is enough for the program's other studies to fall under the same rating—while 92% of Level 2 studies are quasi-experimental. Of the 85 studies, 19 report no usable effect size. Among the remaining 66, the median is +0.15 SD, and weaker designs report higher medians—approximately 0.16 at Level 3 compared with approximately 0.14 at Level 1: less rigor, larger figures. Apps, the largest group in the catalog—20 of 52—cluster between +0.07 and +0.13 SD. Large effects are concentrated in small-group interventions and tutoring, and one-third of the effects of 0.25 SD or more come from assessments aligned with the program (Figure 5).

Less Rigor, Larger Figures: The Evidence for ESSA Catalog by Evidence Level

85 studies and 52 mathematics programs · each effect band is calculated from the studies that report a usable effect size



Source: authors' analysis of the Evidence for ESSA mathematics catalog (Innovamat Research Lab, 2026a). The level rating applies to the program, not the study: this is why there are quasi-experiments at Level 1.

Figure 5. Study designs and distribution of effect sizes in the *Evidence for ESSA* mathematics catalog, by evidence level. Includes all program types in the catalog: mathematics programs, apps and platforms, interventions and tutoring, and teacher training. The percentages on the right are calculated from the 66 studies with a usable effect size. Authors' analysis (Innovamat Research Lab, 2026a).

"What about high-quality apps and digital platforms? Are they a good investment?"

The data suggest that they are, but with some qualifications. Their effects (+0.07–0.13 SD) are similar to those of changing curriculum materials, while their cost is usually much lower. If chosen carefully, they may therefore be efficient investments. The fine print: First, many app studies include only students who used them "with fidelity"—frequent users—not everyone assigned to use them. This tends to inflate the figure because frequent users of a tool are not representative of all students, and analyses that include every student assigned to the tool find much smaller effects. Second, the latest evidence points to the

lower end of the range: the largest recent randomized trial of a platform—*Zearn* in Texas—found +0.07 SD, which was not statistically significant, on the state assessment and +0.11 on an interim assessment (Pane et al., 2025). The chapter's rule also applies here: request the effect for all assigned students, measured using an external assessment.

"Who conducts all these studies?"

Mostly universities and research firms. In the certified catalog, university teams authored 35% of the studies. In particular, the Fuchs lab at Vanderbilt and Clarke's group in Oregon conduct many of the intervention trials, while the CRRE at Johns Hopkins works under contract for publishers. Research firms—LXD, Mathematica, SRI, AIR, RMC, Empirical Education, WestEd, RAND, HumRRO, and others—authored 40%, while publishers' internal teams authored 20%. Only one-third have undergone peer review. On publishers' websites, the proportion is reversed: two out of three documents are authored by the publishers themselves. One notable structural finding is that almost all independent randomized trials of mathematics programs exist because of public funding, particularly funding from the Institute of Education Sciences.

"Does the same thing happen outside the United States?"

The evidence is even scarcer. Outside the United States, there is almost no evidence about specific programs, and research focuses on instructional approaches. However, the conclusions are not very different. In the United Kingdom, *Mathematics Mastery* produced +0.07 SD when the two EEF foundation trials were combined (Jerrim and Vignoles, 2016)—the typical figure for changing curriculum materials. In low- and middle-income countries, "structured pedagogy"—materials combined with training and support—is one of the most cost-effective investments identified by the World Bank panel (GEEAP, 2023). Two recent European trials of online tutoring—about +0.22 SD in Italy (Carlana and La Ferrara, 2025) and +0.26 SD in Spain (Gortazar et al., 2024)—confirm the strength of well-designed tutoring. The effect-scale curve holds across different contexts.

6 | Key Takeaways

Three statements summarize the chapter. First, evidence is the exception. Most mathematics programs have no studies, and very few have studies capable of showing with high confidence that any observed change resulted from using the program. Second, actual effects are modest—about one-tenth of a standard deviation—but they may often represent a good investment if alternative products have similar costs. Finally, the key is not just the curriculum materials, but the complete package that a good program can provide: its effect more than doubles when the materials are accompanied by practice-based training. Chapter 3 explains the underlying reasons by examining the factors that determine learning—from education policy to families, with teachers at the center—and identifying which factors a school can change.

Analyze the evidence for what you want to implement, expect modest improvements, and prioritize curriculum materials with strong implementation support.

Further Reading

The three entry points for readers who want to explore further on their own without being researchers are marked with ▸.

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). Achievement effects of four early elementary school math curricula: Findings for first and second graders (NCEE 2011-4001). Institute of Education Sciences.

Agodini, R., Harris, B., Seftor, N., Remillard, J., & Thomas, M. (2013). After two years, three elementary math curricula outperform a fourth (NCEE 2013-4019). Institute of Education Sciences.

Bailey, D. H., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness, 10(1), 7–39.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. Psychological Science in the Public Interest, 21(2), 55–97.

▸ Bellwether Education Partners (2023). Rounding up: An analysis of math curriculum effectiveness studies. Bellwether.

Blazar, D., Heller, B., Kane, T., Polikoff, M., Staiger, D., et al. (2020). Curriculum reform in the common core era: Evaluating elementary math textbooks across six US states. Journal of Policy Analysis and Management, 39(4), 966–1019.

Carlana, M., & La Ferrara, E. (2025). Apart but connected: Online tutoring, cognitive outcomes, and soft skills. American Economic Review, 115(10), 3487–3513.

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

Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., & Smolkowski, K. (2023). Examining the impact of a first grade whole number intervention by group size. Journal of Research on Educational Effectiveness, 16(2), 326–349.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. Educational Evaluation and Policy Analysis, 25(2), 119–142.

Garet, M. S., et al. (2010, 2011). Middle school mathematics professional development impact study (NCEE 2010-4009; NCEE 2011-4024). Institute of Education Sciences.

Garet, M. S., et al. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher professional development (NCEE 2016-4010). Institute of Education Sciences.

GEEAP — Global Education Evidence Advisory Panel (2020, 2023). Cost-effective approaches to improve global learning. Banco Mundial.

Gortazar, L., Hupkau, C., & Roldán-Monés, A. (2024). Online tutoring works: Experimental evidence from a program with vulnerable children. Journal of Public Economics, 232, 105082.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3), 172–177.

Innovamat Research Lab (2026a). Math program efficacy studies: base de datos abierta de evidencia de eficacia de programas de matemáticas en EE. UU. https://docs.google.com/spreadsheets/d/1c4LJ_Kyt0oQt-inlFzhyEdQ8rkiPOn1jNEfApO2wHB0

Innovamat Research Lab (2026b). Publisher efficacy claims: corpus de los documentos de eficacia publicados en los portales de las seis grandes editoriales de matemáticas de EE. UU. https://docs.google.com/spreadsheets/d/1XACeSy6m91OFSLLSx-IzNNdkmIlGhTTe196IroAMI3M

Institute of Education Sciences (2007). Effectiveness of Cognitive Tutor Algebra One implemented at scale (Award R305A070185). U.S. Department of Education.

Institute of Education Sciences & National Science Foundation (2013). Common guidelines for education research and development. U.S. Department of Education y National Science Foundation.

Jaciw, A. P., Hegseth, W. M., Lin, L., Toby, M., Newman, D., Ma, B., & Zacamy, J. (2016). Assessing impacts of Math in Focus, a "Singapore Math" program. Journal of Research on Educational Effectiveness, 9(4), 473–502.

Jerrim, J., & Vignoles, A. (2016). The link between East Asian "mastery" teaching methods and English children's mathematics skills. Economics of Education Review, 50, 29–44.

Kennedy, M. M. (2016). How does professional development improve teaching? Review of Educational Research, 86(4), 945–980.

Koedel, C., Li, D., Polikoff, M. S., Hardaway, T., & Wrabel, S. L. (2017). Mathematics curriculum effects on student achievement in California. AERA Open, 3(1), 2332858417690511.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, 49(4), 241–253.

‣ Kraft, M. A. (2023). The effect size benchmark that matters most: Education interventions often fail. Educational Researcher, 52(3), 183–187.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher *coaching* on instruction and achievement: A meta-analysis of the causal evidence. Review of Educational Research, 88(4), 547–588.

Kraft, M. A., Schueler, B. E., & Falken, G. T. (2026). What impacts should we expect from tutoring at scale? Exploring meta-analytic generalizability. Review of Educational Research. Publicación anticipada en línea.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? Educational Researcher, 48(3), 158–166.

‣ Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. Educational Evaluation and Policy Analysis, 41(3), 260–293.

Lynch, K., Gonzalez, K. E., Hill, H. C., & Merritt, R. (2025). A meta-analysis of the experimental evidence linking mathematics and science professional development interventions to teacher knowledge, classroom instruction, and student achievement. AERA Open, 11, artículo 23328584251335302.

Nickow, A. J., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for preK–12 learning: A systematic review and meta-analysis of the experimental evidence. American Educational Research Journal, 61(1), 74–107.

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

Pane, J. F., Doss, C. J., Todd, I., & Seaman, D. (2025). Efficacy of Zearn Math over two years in grades 3 to 5: An experiment in Texas (EdWorkingPaper 25-1211). Annenberg Institute, Brown University.

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. AERA Open, 7(1), 1–29.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & 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.

Saxe, G. B., Gearhart, M., & Nasir, N. S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. Journal of Mathematics Teacher Education, 4(1), 55–79.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. Review of Educational Research, 78(3), 427–515.

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating Math Recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. American Educational Research Journal, 50(2), 397–428.

Whitehurst, G. J. (2009). Don't forget curriculum. Brookings Institution.

Funding and declaration of interests. *This publication was funded by Innovamat Education and is part of a science communication series on education promoted by the organization. Carles Aulés and Marc Colomer are members of Innovamat. Affiliations and funding are disclosed as part of the publication's commitment to transparency. The authors are responsible for the content and conclusions.*

Acknowledgments. *We would like to thank Ángela Pérez for reviewing this chapter and helping improve its quality. We would also like to thank Marta Soler for her layout work.*