Innovamat

Does It Work? How to Assess the Evidence Behind a Mathematics Program

A guide to interpreting evidence on math program effectiveness

Marc Colomer,
Director of Research at the Innovamat Research Lab

Carles Aulés,
Research Specialist at the Innovamat Research Lab

Caterina Calsamiglia,
ICREA Research Professor and Head of the Computational Social Science and Humanities unit at the Barcelona Supercomputing Center (BSC)

Innovamat Research Series

Summary

When a school adopts a new math program, a common question is “Does it work?” This report explains how to evaluate an educational program and interpret an impact study. It focuses on three main ideas:

•We cannot assess whether an educational program works in the abstract. The question we can ask is: “What difference does using this program make compared with the available alternative, for a specific outcome, population, and context?”

- Not all evidence carries the same weight or is equally informative. The rigor of an impact study depends on its ability to estimate what would have happened to those same students without the program. This requires constructing a fair comparison with comparable students who did not receive it.

- A good impact evaluation also requires a sound theory of change: What should change, through what mechanism, and under what conditions? We need to specify how the program is expected to bring about change and recognize implementation as an essential part of that process.

The report concludes with eight questions to guide readers through any impact study, followed by answers to common questions about reports on program effectiveness.

When a school adopts a new math program[1], teachers and families often ask: "All right, but does it work?"

It is a reasonable question. There is a growing need to assess whether educational practices have an impact and whether that impact is positive or negative. In medicine, we expect treatments to be evaluated before they are used. That expectation has yet to become firmly established in education. Fortunately, assessing the effectiveness of the practices and resources introduced into schools and making evidence-based decisions are becoming increasingly common.

This growing demand for evidence also presents a challenge. In some cases, the word "evidence" has become a sales pitch. Materials arrive at schools accompanied by studies, percentages, and seals of approval claiming proven improvements in learning. The growing demand for empirical evidence is welcome. However, we need to know what question that evidence answers and how useful the answer is to us.

To begin with, we cannot determine whether a program works in the abstract. We can, however, answer a more specific question: **What difference does using this program make compared with the available alternative, for a specific outcome, population, and context?**

How we frame the question matters. A program's effect depends on what we compare it with, what outcome we measure, who participates in the study, and the conditions under which it is implemented. A positive result tells us what happened there; determining whether we can expect something similar here requires much more.

This report explains how to construct a comparison that allows us to estimate a program's effect and how randomization and other evaluation designs help us do so. It also explores what it means to measure improvement in mathematics, how to interpret an effect's size and precision, and why a program may work in some settings but not others. The aim is to help us interpret an impact estimate by asking **what difference it captures, how it was estimated, and how far we can expect a similar result in our own context.**

By the End of This Report,

you will know what to ask when someone tells you that a math program "works": what it was compared with, what students actually learned, how much they improved, and under what conditions that result was achieved. Above all, you will be able to assess what that evidence tells you about what may happen in your classroom.

[1] We define a math program as a package of educational resources that combines materials (such as textbooks, teacher guides, and hands-on learning materials) with implementation support (such as professional development and coaching).

1 | Evaluation Requires a Fair Comparison

Imagine that a school launches a new math program. After one year, its students' results have improved substantially. **Can we say that the program improved their performance?**

Not on that evidence alone. The results may have changed for many reasons. Perhaps a new teacher arrived, the student cohort is different, more time was devoted to mathematics, or the teaching team received additional training. **Observing an improvement tells us that something changed; figuring out what caused it is another matter.**

This problem is not unique to education. Imagine that a pharmaceutical company advertises its new cough syrup by saying that 95% of the people who took it recovered from their colds within a week. Is that enough to conclude that the syrup works? Many people recover from a cold within a week without taking anything. To know whether the syrup had any effect, we need to know what would have happened to those same people if they had not taken it.

Consider another example. Imagine that a study concludes that people who use more sunscreen are also more likely to get sunburned. Does sunscreen cause sunburn? People who use more sunscreen may also spend much more time in the sun, which could explain the association.

Observing that X and Y occur together does not necessarily mean that X caused Y. In research, this distinction between association and causation is fundamental and is summarized in a familiar phrase: **correlation does not imply causation**.

A good effectiveness study seeks to establish whether doing *X*—for example, using a math program—caused a change in *Y*—for example, student learning. Simply observing that the two occur together is not enough.

Let's return to the school. Suppose the new program is introduced at the same time that a highly experienced and especially motivated math teacher arrives. At the end of the school year, the results improve. **Did the program work? Was it the teacher? Did both contribute?**

It is impossible to know from the data for that one school year. Too many things changed between "before" and "after." The comparison we need is different: **How would those same students have performed if they had not received the new program?**

This alternative scenario is what we call the **counterfactual** (Gertler et al., 2016), and it presents a fundamental difficulty: We can never observe it directly because the same student cannot receive the program and not receive it at the same time.

Counterfactual. What would have happened to these same students if they had not received the program or intervention. We cannot observe it directly, so the study's credibility depends on how well it estimates this alternative outcome.

How Can We Approximate Something That Did Not Happen?

By making a comparison. We look for a group of students who did not receive the program but are as similar as possible to those who did. What happens to them helps us estimate what would have happened to the first group without the program. This is the **comparison or control group**.[2]

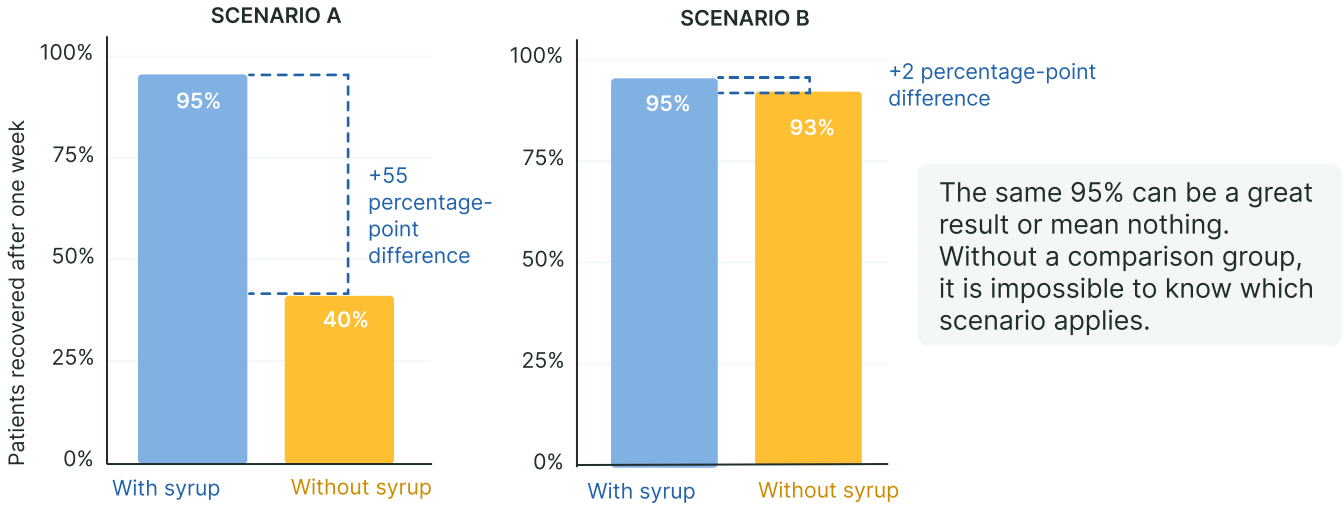
[2] We use the term *control* when discussing experimental studies, whereas *comparison* is more appropriate for quasi-experimental or correlational studies.

Let's return to the syrup example. If 95% of those who take the medicine recover within a week, that figure alone tells us very little. If 93% of comparable people who do not take it recover, the syrup is probably doing little. If only 40% recover, we would have a much clearer indication that the medicine is having an effect. The 95% figure becomes informative when we have something to compare it with.

The key question then becomes **how we construct a sufficiently fair comparison**. When we can randomly assign who receives the program, the two groups are comparable on average, even in characteristics we do not know how to measure. When random assignment is not possible, we need other designs that construct an equally credible comparison. This is the logic behind **randomized experiments** and **quasi-experimental designs**.

> "Good evaluation is not just about measuring change;
> it is about constructing a fair comparison."

The Comparison Gives Us Answers



The same 95% can be a great result or mean nothing. Without a comparison group, it is impossible to know which scenario applies.

Figure 1. An illustration of the syrup study with a comparison group in two scenarios. In scenario A, it is uncommon for patients to recover without taking the syrup. In scenario B, however, almost all patients (93%) recover naturally without taking anything. Without a fair comparison, we cannot interpret the 95% figure.

Higher scores after adopting a new math program do not by themselves show that the program works because other changes during the same period may help explain the improvement. Evaluating causal impact means estimating what would have happened to those same students without the program. This requires constructing a fair comparison with comparable students who did not receive it.

2 | How Do We Construct a Fair Comparison?

Chance as a Tool

The simplest way to construct one is to use chance. Imagine that we have a group of schools interested in trying a new math program. Instead of letting each school decide whether to use it, we randomly select which schools will receive it and which will continue with the usual alternative. Notice what is being randomized: random assignment determines who receives the program and who serves as a comparison among schools that had already agreed to participate.

If we have enough schools, chance makes the two groups comparable on average before the study begins. Some schools will have more experienced teachers and others less experienced teachers; some will have students with higher initial scores and others with lower scores; some will be more motivated and others less motivated. Because assignment was random, none of these characteristics should be systematically concentrated in either group.

This includes characteristics that we do not even know how to measure. We can compare schools by their students' socioeconomic status, prior achievement, or school size, but there will always be other differences that are harder to observe, such as the teaching team's expectations, family involvement, or school culture. **The key advantage of random assignment is that it balances them on average without our needing to know what they are.** Therefore, if students who received the program subsequently perform better than those in the comparison group, we have a much stronger reason to attribute that difference to the program.

> Randomization makes assignment to the program independent of schools' preexisting characteristics, creating groups that are comparable on average. Individual schools will still differ, but the subsequent difference in outcomes can nevertheless be interpreted as a causal effect with much greater confidence.

This is the principle of a **randomized experiment**, or *randomized controlled trial* (RCT). It uses the same logic as a clinical trial, in which some people receive the treatment and others receive a placebo or the usual alternative, and the difference between the two groups allows researchers to estimate the treatment's effect. In education, however, this approach has some distinctive features.

Educational Experiments Have Their Own Challenges

We often cannot randomly assign individual students. If the program changes how teachers teach, the usual approach is to assign **entire classrooms or schools**, because otherwise the same teacher would have to use different methods with students in the same class, and what happens with some students may affect the others.

Moreover, unlike in a medical trial, everyone usually knows which program they are using. Teachers know that they are teaching with new materials, and students may know that they are participating in an intervention. Therefore, some of what we observe may result from novelty, increased attention, or the additional effort accompanying any experiment.

There is one final important difference. **The effects of a curriculum resource depend on more than** the resource itself. What happens in the classroom depends on how teachers use it, how students respond, and conditions at the school. Therefore, two schools may receive exactly the same program and implement it in very different ways. When reading the results of an experiment, remember exactly what was evaluated. The study tells us **the difference that the program made, relative to the comparison alternative, when implemented in that way, with that population, and in that** context.

What If We Cannot Use Chance?

We cannot always assign a program randomly. Perhaps all the schools want to use the program and do not want to wait, perhaps the education authority has already implemented it in certain schools, or perhaps ethical or practical reasons make random assignment impossible.

In these cases, researchers construct the comparison in other ways using what are called **quasi-experimental designs**. The logic remains exactly the same: find a group that allows us to approximate as closely as possible what would have happened without the program.

Imagine that an education authority decides to offer a program first to schools with test scores below a specified threshold. Researchers may compare schools just above and just below the threshold if they have reason to believe that the schools are sufficiently similar and that falling on one side or the other of the cutoff determines participation. Other studies use policy changes, phased implementation, or administrative rules that create comparable groups.

A good quasi-experimental design can provide very strong causal evidence. When the rule used to form the groups provides a credible basis for comparison and the assumptions required by the design are explicitly tested, the resulting evidence can be almost as convincing as evidence from a randomized experiment. When reading such a study, we need to focus on a different set of questions. Without the protection provided by random assignment, we must examine more carefully **why we can consider the two groups comparable and what assumptions the design requires**.

The Danger of Comparing Self-Selected Groups

Let's return to the example of schools volunteering to try a program. Suppose that after one year their results improved more than those of schools that did not participate. Did the program work?

Perhaps it did, but something else may also have happened. Schools that volunteered may have especially motivated teachers, leadership teams that place great importance on mathematics, or a greater capacity to introduce changes. The opposite may also be true: they may be schools with particularly low scores that were actively seeking a solution.

If those characteristics also affect the results, the difference we observe may reflect **who decided to participate**. This is what we call **selection bias**, and it is especially difficult to detect because some relevant differences may not appear in the data. We can make the two groups very similar in size, governance, socioeconomic status, or initial results, yet they may still differ in an important characteristic that we are not measuring.

> Comparing those who received a program with those who did not may seem straightforward, but it can be highly misleading. If the people, schools, or teachers who choose to participate differ from those who do not, those differences may be mistaken for the program's effect.

The Quality of the Comparison Determines the Strength of the Evidence

Therefore, when we read a study, one of our first questions should be **how participants were assigned to the program** or the comparison group. Random assignment gives us a very strong foundation; any other rule requires us to understand why it allows us to consider the groups comparable.

This explains why studies that use the same word, "evidence," support very different levels of confidence. A study may find that students using a program perform better; if those students were selected in a way that makes them systematically different from the comparison group, we still do not know how much of that difference is due to the program. Another study may use a much more rigorous comparison and find a smaller difference. Paradoxically, **the second result may be much more informative**. Assessing the credibility of the comparison must therefore come before considering the effect size.

One Way to Classify Evidence

This difference among designs is also why some systems classify evidence according to the confidence it provides in a causal claim. One of the best-known frameworks comes from the **Every Student Succeeds Act (ESSA, 2015)** in the United States, which distinguishes four levels of evidence: “strong,” “moderate,” “promising,” and “demonstrates a rationale”. The higher levels require designs that allow for more convincing causal inference. The lowest level includes programs supported by a research-based rationale that have not yet been sufficiently evaluated.

The Four Tiers of the ESSA Evidence Scale

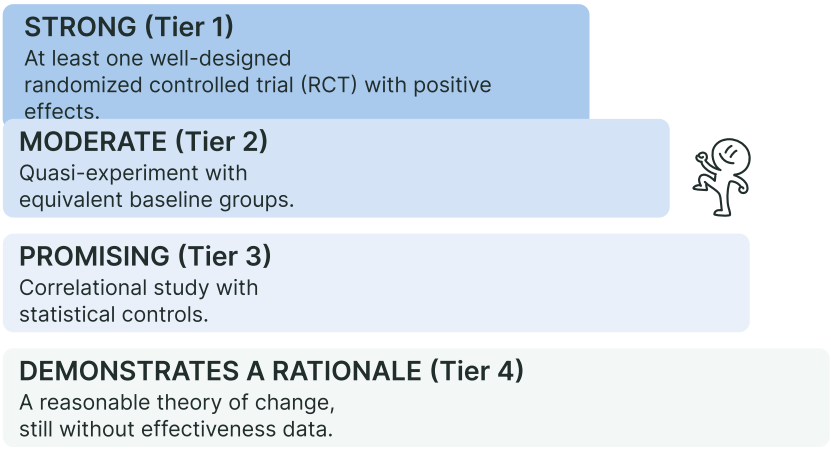


Figure 2. The four evidence levels of the Every Student Succeeds Act (2015), ranked by the strength of support for a causal claim.

Interpret this carefully. **This scale ranks the strength of the available evidence about programs, not the quality of the programs themselves.** A program at a low level may be promising and warrant a more rigorous evaluation. A program with strong evidence is supported by studies that allow us to attribute the observed results to the program with much greater confidence.

Even a good experiment does not answer every question. It tells us whether we found a difference between two comparable groups, but we must still ask **what was measured, how much students improved, and whether we can expect something similar in our own context**. We will address these questions next.

The fundamental question is why we should believe that the observed difference is due to the program. The answer depends primarily on how the comparison was constructed.

3 | Before Measuring Impact, We Need to Know What We Expect to Change

We now know how to construct a comparison that allows us to estimate a program's effect. However, we still need to ask a more basic question: **What do we expect to change if the program works?**

Imagine a new math program that encourages students to work on more open-ended problems. If it works, students will have more opportunities to explain how they reached a solution. This may improve their reasoning skills, and over time, that improved reasoning may lead to better performance on unfamiliar problems.

This chain of steps is known as a **theory of** change: an explanation of how we expect an intervention to produce its results. It does not need to be complicated. At its core is a simple question: **Why do we think doing A should produce a change in B?**

From the Program to Learning

Let's continue with the same example. For a program that asks teachers to use more open-ended problems, the expected sequence would have four stages.

1. The program is implemented. Teachers receive training and begin using the new activities.

2. Classroom practice changes. Students actually work on more open-ended problems and have more opportunities to explain their strategies.

3. Learning changes. Students develop better reasoning and problem-solving strategies.

4. The outcomes that interest us change. Students perform better on a problem-solving test or a broader measure of mathematical proficiency.

Each arrow in that chain represents a hypothesis that may not hold. Perhaps the program increases the number of problems students work on but leaves their strategies unchanged. Perhaps students develop new strategies but do not know how to use them when they encounter problems different from those they practiced. Perhaps they improve in problem solving, but the change is too small to be observed on a general math test.

Therefore, when a study finds no effect, we have a more useful question than "Does the program work or not?" We can ask **where in the chain the expected change** failed to occur. Teachers may not have been able to implement the program as intended. The activities may have been used without truly changing students' learning experience. Learning may have changed, but the test used may not detect it. Or the initial hypothesis may simply have been incorrect, and working on more open-ended problems may not produce the type of learning we expected.

Examining the entire chain helps us distinguish among these possibilities. It also changes how we interpret a positive result. If we find an improvement on the final test, we can check whether we also observe the expected intermediate changes. When they appear, the proposed explanation is more coherent. When they are absent, we should interpret the result more cautiously or revise our explanation of how the program works.

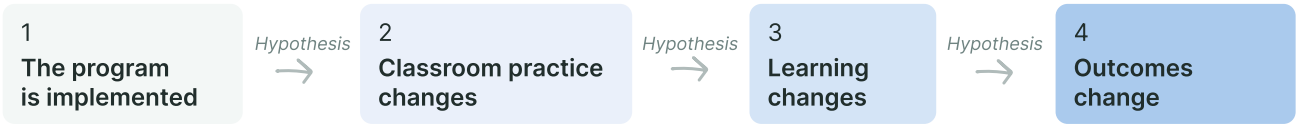


Figure 3. A theory of change connects the intervention to the outcomes we expect to observe. Each step in the chain represents a testable hypothesis.

A theory of change explains why we expect a program to produce specific outcomes. It allows us to move from “this program works” to a much more useful question: What should change, through what mechanism, and under what conditions?

What Does “Improving in Mathematics” Mean?

The theory of change also requires us to specify what we mean by learning. A program may be designed to improve calculation skills, conceptual understanding, problem solving, or reasoning. It may also seek to change students’ confidence, their attitudes toward mathematics, or their willingness to tackle difficult problems. All these outcomes may be relevant, and **we will know whether a program achieved its goal only if we measure the appropriate outcome**.

Therefore, we should distinguish among different types of outcomes.

- **Proximal outcomes.** Those most directly related to what was taught, such as mastering a specific strategy or correctly solving the types of exercises practiced.

- **Transfer.** Whether students can use what they learned in new situations involving different problems, different wording, or tasks they have not practiced directly.

- **Broad or lasting outcomes.** More general changes or changes that persist over time, such as an overall improvement in mathematical proficiency, greater confidence, or later effects on students’ educational trajectories.

All three types of outcomes may be important, but they answer different questions. A program may produce a clear effect on a proximal outcome and a much smaller effect on a more general test. This may mean that learning has not yet transferred, that the intervention targets a very specific skill, or simply that we are measuring something different from what the program is intended to change.

The Test Is Part of the Question

Choosing a test is therefore much more than a technical detail to be addressed after the study has been designed. **The test largely determines what claim we can make about the program.**

Imagine that a program teaches a specific problem-solving strategy and that, at the end of the school year, students perform better on exercises that are nearly identical to those they practiced. We can conclude that they learned that strategy. To conclude that their general problem-solving ability has improved, we would need to examine how they perform on unfamiliar tasks.

Test selection matters for another reason. Tests designed by program developers are usually especially well aligned with what is taught, which tends to make the effect appear larger. Cheung and Slavin (2016) reviewed 645 educational studies and found that effects measured using tests created by the researchers themselves were approximately twice as large as those measured using independent tests. Meta-analyses of STEM interventions find the same pattern (Lynch et al., 2019).

This makes sense. If we spend months practicing a particular type of exercise and then assess precisely that type of exercise, the assessment closely reflects what students have practiced. A broader, independently developed test sets a higher bar: **Can students use what they learned when the task changes?**

None of this makes a program-specific test a poor tool. If we want to know whether students learned the exact content or strategy that the program aims to teach, a program-specific test may be the best available option. **Each test supports a different claim**, so we should always check who designed the test and how closely it was aligned with the program.

QUESTIONS WE SHOULD ASK

- What was the program intended to change?
- What mechanism should produce that change?
- What outcome was measured?
- Does the test measure what we actually want to improve?
- Is the test closely related to what students practiced, or does it require them to transfer what they learned?

To interpret an impact study, we need to understand what change the program was intended to produce, what mechanism was expected to lead to that outcome, and which parts of that chain we can observe. The theory of change also tells us where to look when we want to know whether we can expect a similar result in another school, an issue we will revisit later.

4 | How Large Is the Effect, and How Precise Is the Estimate?

We have seen how to construct a comparison that allows us to attribute a difference to the program and how to decide what outcome we want to measure. Now suppose that we conduct the study and find that students who used the program performed better. The next question is **how much better**.

A study's result is always an estimate. To interpret it, we need to answer two questions: What is the magnitude of the difference, and how precisely did we estimate it? The effect size answers the first question; the **confidence** interval addresses the second. Studies also commonly report a **p-value**, which helps us assess how compatible the data are with the absence of an effect but tells us nothing about whether the observed difference is large or educationally meaningful.

Effect size. The size of the difference between the group that used the program and the comparison group, measured in standard deviations (SD). It is usually expressed as Cohen's d or Hedges' *g*, two closely related measures named after the statisticians who developed them.

Not all tests use the same scale. One may be scored from 0 to 10, another from 0 to 100, and another may use its own scale. To compare results across studies, we express the difference in relation to the variability in results among students, as measured by the standard deviation. An effect of 0.20 SD means that the difference between the two groups is equivalent to one-fifth of that typical variability.

In other words, if we line up the 25 students in a class by their test scores, an effect of 0.20 SD would move the student in the middle up about two places.

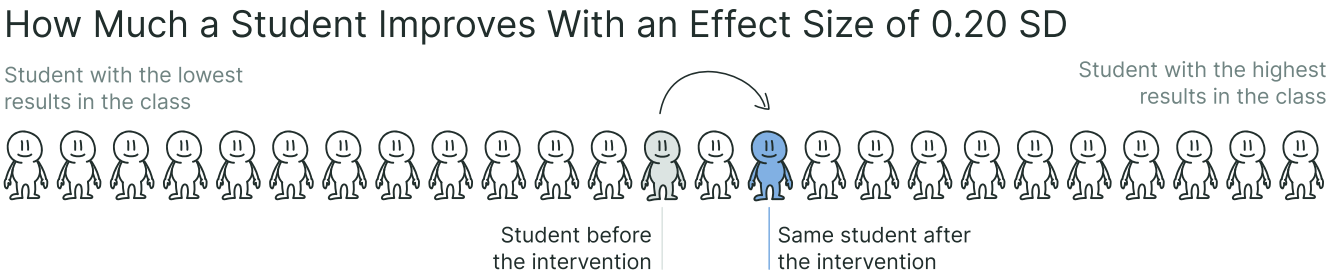


Figure 4. The 25 students in a class ranked by their test scores. An effect of 0.20 SD moves the student in the middle up about two places.

An estimated effect of 0.20 SD is still subject to uncertainty. If we repeated the study with other comparable schools, we would probably obtain a somewhat different figure. The confidence interval expresses this uncertainty. An estimated effect of 0.20 SD with an interval from 0.15 to 0.25 is a fairly precise estimate. The same estimated effect of 0.20 SD with an interval from −0.05 to 0.45 is much more uncertain because the data are compatible with both a small negative effect and a large positive one. The problem is more common than it may seem: a review of large-scale educational trials funded in the United States and the United Kingdom found that many produce estimates that are too imprecise to be informative (Lortie-Forgues & Inglis, 2019). Therefore, always read the effect size together with its confidence interval.

An effect size should be interpreted together with its confidence interval. We need to know both the estimated magnitude and the uncertainty surrounding it.

Is 0.20 SD Large or Small? It Depends on What We Compare It With

To interpret an effect size, we need a reference point: the effects typically found in similar studies.

Kraft (2020) proposed benchmarks suited to school settings, based on the distribution of effects obtained in randomized educational trials, and updated them three years later using a much larger sample. He compiled 3,426 effect sizes from 973 randomized trials and found that effects tend to be modest. More than one-third were below 0.05 SD, the median was 0.10 SD, and only 10% exceeded 0.57 SD (Kraft, 2023). Based on this distribution, he proposed a scale specific to educational interventions: less than 0.05 SD is a small effect, between 0.05 and 0.20 SD is a medium effect, and 0.20 SD or more is a large effect. A very similar distribution appeared in 138 trials conducted in low- and middle-income countries, also with a median of 0.10 SD (Evans & Yuan, 2022).

The choice of benchmark matters. Other frameworks have also been used in education, such as Cohen's (1988), developed for the behavioral sciences in general, and Hattie's well-known 0.40 threshold (2009). Both have been highly influential and draw on a wide range of sources, rather than solely on effects from rigorous experiments in real educational settings. Applying them directly would lead us to dismiss effects as small even when they are meaningful relative to those typically achieved by educational interventions.

Therefore, an effect of 0.20 SD may seem modest in absolute terms while still being an important educational result. In addition to asking how large the effect is, we should ask another question: **How does it compare with what similar interventions achieve?**

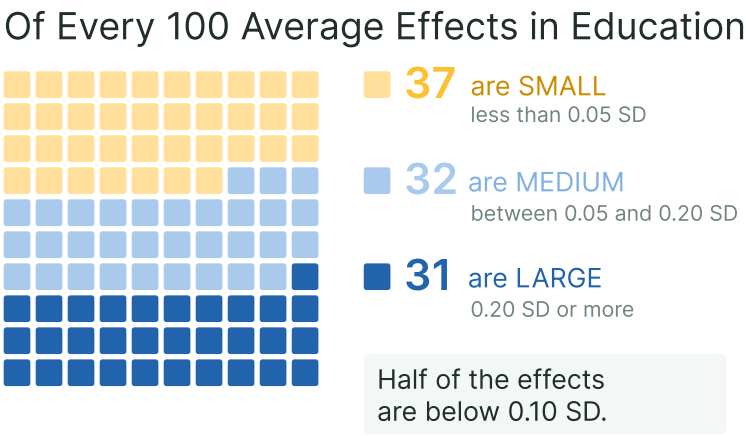


Figure 5. Distribution of 3,426 effect sizes from 973 randomized educational trials, showing the median and the three ranges proposed by Kraft. Authors' analysis based on Kraft (2020, 2023).

There is a second important caution. The same effect size may represent a more or less meaningful improvement depending on students' grade level.

> "In education, an effect of 0.20 may already be an important result. Interpreting it also requires considering how much progress students typically make at that age."

Hill, Bloom, Black, and Lipsey (2008) estimated how much progress students typically make on standardized tests over a school year and showed that this progress varies greatly by age. In the early elementary grades, observed growth may approach one full standard deviation per year. By fifth or sixth grade, it is around 0.40 SD, and in the final years of secondary school, it falls below 0.25 SD. In other words, there is considerable room for improvement in the early grades, and as students become more proficient, demonstrating progress becomes increasingly difficult.

Typical annual progress therefore provides a second reference point for interpreting effect sizes, alongside the effects of other interventions.

A Student's Potential for Improvement as They Progress

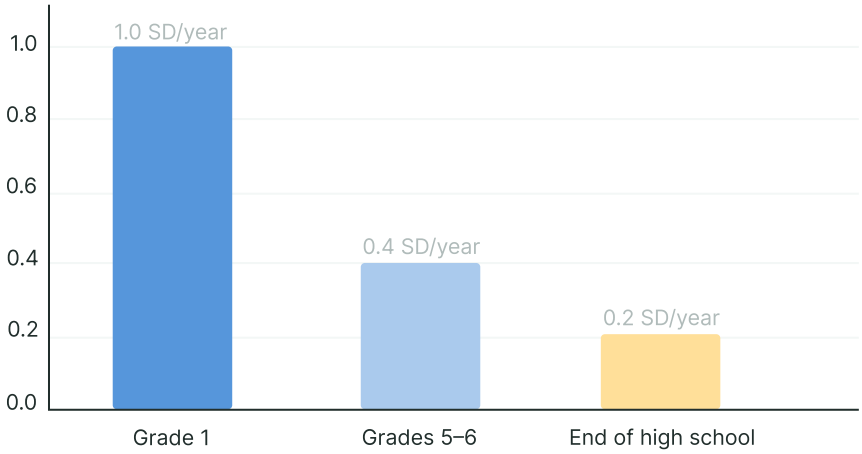


Figure 6. Typical annual growth by grade, based on Hill et al. (2008).

An effect of 0.10 SD, a typical effect size for math curriculum programs (Slavin & Lake, 2008; Pellegrini et al., 2021), represents approximately one month of the progress typically observed during a school year in the early elementary grades. In the final years of secondary school, where observed annual growth is much lower, the same 0.10 SD represents a considerably larger proportion of one school year's progress.

These comparisons are sometimes expressed as "months of learning." They can make an effect size easier to understand, but they should be interpreted very cautiously. A study observes a difference in results and compares it with typical growth over a school year; expressing it in months is a separate conversion that depends on the grade, test, and reference used.

Months It Takes a Student to Improve by 0.10 SD

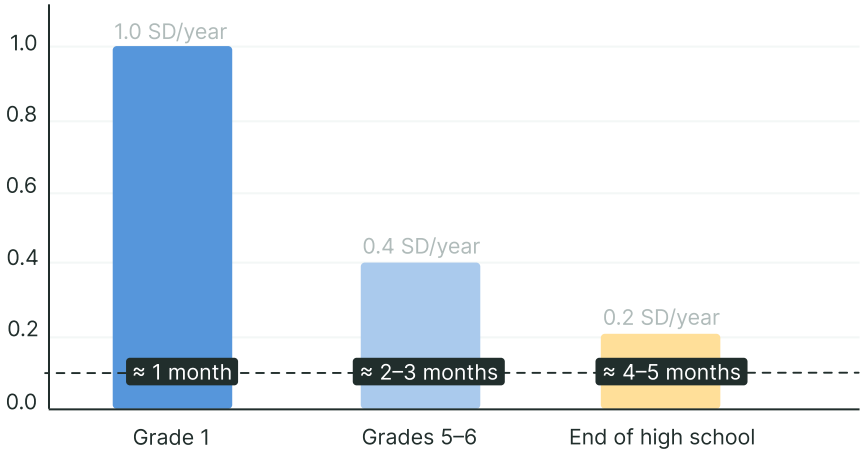


Figure 7. Approximate number of months of typical progress represented by an effect of 0.10 SD, by grade, based on Hill et al. (2008). These equivalences are approximate because impact studies do not directly measure months of learning.

A study's result is an estimate. To interpret it, we need to know **how much the outcome changes**—the effect size—and **how precisely we estimated it**—the confidence interval. We then need to put that effect size in context by comparing it with the effects typically achieved by similar interventions and with students' usual progress at that age. **The same figure may seem small or large depending on the benchmark we use.**

5 | What Does a Study Tell Us About a Program?

So far, we have discussed effects of 0.10 or 0.20 SD as though they were fixed characteristics of the program. We should avoid that interpretation. A study estimates what happened when a specific program was implemented in a particular way with particular students, compared with a specific alternative and using a specific measure. If these conditions change, the effect we observe may also change. In fact, a meta-analysis of 191 experiments on mathematics interventions found that differences in context and design explained much of the variation in results (Williams et al., 2022).

To understand what a study really tells us, we need to consider all the factors discussed so far, because each contributes to the quality of the evidence. A large effect is unconvincing if the groups were not comparable. A very well-designed study may answer a question that is not particularly relevant to us if it measures only tasks closely aligned with the program. A positive effect may also remain uncertain if the confidence interval is very wide.

Even when everything aligns, two rigorous studies of the same program may find different effects without either being wrong. The following questions help explain why.

What Was Used to Measure It?

As we saw in Section 3, the test used partly determines what we mean by "working." Studies of educational interventions typically find larger effects on tests designed for or closely aligned with the program than on more independent measures (Cheung & Slavin, 2016; Lynch et al., 2019). A proximal test may show whether students learned what the program taught directly, while a broader, more distal measure may show whether that learning transfers to new tasks. The important question is **what** the observed effect tells us about students' learning.

What Was It Compared With?

The effect size also depends on the alternative with which the program is compared. Comparing an intervention with loosely structured teaching is very different from comparing it with another well-implemented math program. Both studies may be well designed, but they answer different questions. Therefore, we need to know what the comparison group actually received. **The better the available alternative, the harder it will be to demonstrate an additional benefit.**

When Was It Measured?

The timing of measurement also matters. A program may produce an improvement that is evident at the end of the intervention but then diminishes, disappears, or, in some cases, grows over time. The most comprehensive review of this issue concludes that partial loss of the effect is typical and that follow-up studies find that many interventions retain only a fraction of their initial effect (Bailey et al., 2020). If we want to know whether learning lasts, we need follow-up assessments. An effect observed immediately after the program ends tells us that there was a difference at that time; how long it will last is a separate question.

Under What Conditions Was It Implemented?

Finally, we need to understand what was required to obtain the result. A program evaluated in a few schools with especially motivated teachers, intensive training, and ongoing support from the development team may produce different results when scaled up to hundreds of schools under ordinary conditions. This is a well-documented pattern. In instructional coaching, for example, the effects of large programs are only a fraction of those obtained by small pilot programs (Kraft, Blazar, & Hogan, 2018).

This does not make a small study a poor study. It tells us under what conditions the effect was observed and prompts us to ask which of them were necessary to produce it. As we saw when discussing the theory of change, implementation is one of the links connecting the program to learning, and a good program may have no effect if it does not consistently transform what happens in the classroom (Durlak & DuPre, 2008). Combining high-quality materials, teacher training, and practice-focused support tends to work better than implementing these elements separately (Lynch et al., 2019; Piper et al., 2018). Evidence showing that effects are sustained when a program moves from highly controlled conditions to widespread use under everyday conditions is therefore especially informative.

This is also why the National Research Council committee that reviewed the evaluation of curriculum programs in mathematics concluded that no single study is sufficient and proposed combining three complementary perspectives: impact on outcomes, classroom implementation, and the mathematical quality of the content (Confrey & Stohl, 2004).

Who Conducted the Study?

It is also useful to know who conducted and funded the evaluation. Reviews of education research find that studies conducted or commissioned by those who develop an intervention tend to report larger effects than independent evaluations (Wolf et al., 2020). Small studies also tend to report larger effects than large studies, and published studies tend to report larger effects than unpublished ones (Cheung & Slavin, 2016; Slavin & Smith, 2009; Franco, Malhotra, & Simonovits, 2014).

None of this invalidates a developer-funded study, whose design should be assessed using the same criteria as any other study. These patterns are another reason to value results that have been replicated by independent teams.

What Does This Result Mean for Our School?

A rigorous study tells us what happened under specific conditions. Applying that result to another school requires us to do more than look for studies conducted in schools identical to ours. The important question is **whether the conditions necessary for the program to produce the change are present in our context**.

This brings us back to the theory of change. If the program required specific teacher training, a particular amount of classroom time, or certain teaching practices to improve learning, we need to ask whether those conditions can be replicated in our school. The more the result depends on conditions we cannot replicate, the less informative that effect will be for predicting what would happen here.

For the same reason, rather than looking for “the study” that proves a program works, it is useful to consider **the full body of available evidence**. A single study may provide highly compelling causal evidence about what happened in a specific context. Our confidence increases when we find consistent results across several rigorous studies and when those results hold across different populations, schools, measures, or implementation conditions. Replication is rare in education research, which is precisely why it is so valuable when it occurs (Makel & Plucker, 2014).

If an effect is found only in a small study with extensive support, using a closely aligned assessment immediately after the intervention, we know something important but limited. If it also appears on broader measures, compared with effective alternatives, when the program is implemented at scale, and in different contexts, we have more reason to think the result does not depend on a highly specific set of circumstances.

A study provides an estimate of what happened under specific conditions. To interpret it, we need to consider the quality of the comparison, what was measured, the magnitude and precision of the effect, what the program was compared with, when the results were measured, and the conditions under which it was implemented. The final question is **what we can conclude from the study and why we might expect a similar result in our context**. The more consistent the evidence is across rigorous studies, measures, populations, and different conditions, the more confident we can be that the result does not depend on a particular set of circumstances.

6 | A Guide to Reading the Evidence

Throughout this report, we have seen that deciding whether a program "works" requires much more than finding a study with positive results. **These eight questions provide a practical guide to reading any study of an educational program.**

1. What is being compared? Identify what students using the program receive and what the comparison group receives. The effect is always a difference relative to a specific alternative.

2. Can we attribute that difference to the program? Check whether participants were randomly assigned and, if not, what strategy supports the conclusion that the groups were comparable before the study began.

3. What was the program intended to change, and how? Examine its theory of change and identify which intermediate steps can be observed.

4. What was actually measured? Ask whether the assessment captures what the program taught directly or requires students to transfer what they learned to new tasks, and whether it measures the outcome that interests us.

5. How large is the effect, and how precise is the estimate? Read the effect size alongside its confidence interval. Compare the estimated effect with those of similar interventions and with students' typical progress at that age.

6. Under what conditions was the result obtained? Check what the program was compared with, when outcomes were measured, and what training, time, and support schools had available.

7. What do we know beyond this study? Look for other rigorous, independent studies of the program and check whether they find consistent results across different measures, populations, and schools.

8. What does this evidence mean for our context? Consider whether we can replicate the conditions that appear necessary for the program to produce the change.

Ask what the study demonstrates, how confident we can be in its findings, and how relevant the evidence is to the decision we face.

7 | Frequently Asked Questions

"If all studies have limitations, how does the evidence help me?"

Evidence does not replace professional judgment; it makes that judgment better informed. Decisions must also take into account students' needs, available resources, cost, and the school's capacity to implement the program. A moderate effect may be highly valuable if the program is affordable and easy to implement, but far less compelling if it requires a disproportionate investment. Evidence can help us set realistic expectations, compare alternatives, identify implausible claims, and show us which conditions appear necessary for a program to produce particular results.

"We have used the same program for three years, and the results have clearly improved. Doesn't that count?"

It counts, and it is valuable information about what is happening at your school. The difficult part is determining how much of that improvement was caused by the program. Over those three years, the teaching staff, teachers' training, the time devoted to mathematics, or the characteristics of the students may have changed. It is also useful to consider whether the improvement is specific to your school or part of a broader trend in the area. **"Our results have improved since we began using the program" and "the program caused that improvement" are different claims**, and the first may be entirely true long before we have enough evidence to support the second.

"My program provider shows me a study with an effect of 0.45 SD. Should I be pleased or skeptical?"

There is reason for both enthusiasm and scrutiny. An effect of 0.45 SD is more than four times the median effect size in educational trials and more than twice the threshold at which Kraft considers an effect large. If the estimate is credible, it would be an exceptional result. That is precisely why it is worth understanding where the figure comes from. What assessment was used? What alternative was it compared with? What is the confidence interval? Were outcomes measured again later? Was the program evaluated in a few schools with extensive support or under conditions similar to typical implementation? Have other researchers found similar results? **An exceptional effect is a good reason to read the study particularly carefully.**

"How much should I trust an evidence certification?"

A certification indicates that someone has reviewed the study against predefined criteria, which is very useful. It is similar to a vehicle inspection certificate, which confirms that a car has passed a rigorous inspection without guaranteeing that it will perform the same way on every road. It tells us about the available evidence and its quality, but we still need to ask what was measured, what the program was compared with, and under what conditions the result was obtained. The reverse is also true. **A program without certification may simply be used in a setting where no organization** offers such certification, or no one may have submitted its studies for review. In that case, we will need to examine the studies directly and apply the questions in this report.

"Who reviews studies and can help me determine whether the evidence is strong?"

Specialized organizations review educational studies using common criteria and synthesize the findings. In the United States, the What Works Clearinghouse reviews the methodological quality of studies (What Works Clearinghouse, 2022), and Evidence for ESSA organizes evidence about programs according to the levels established by law. EdReports assesses curriculum quality and alignment with standards without assessing impact. In Europe, the closest equivalent is the United Kingdom's Education Endowment Foundation, which funds rigorous evaluations and synthesizes educational evidence.

We can also consult studies published in peer-reviewed journals, where other experts examine the methods, analyses, and conclusions before publication. Peer review provides some assurance that the results have been interpreted correctly and that the limitations have been clearly disclosed. However, a study's causal strength still depends on how the comparison was constructed and which question the design can answer. Evidence-synthesis organizations, certifications, and peer review help guide us, but none replaces the fundamental questions in this report.

8 | In Closing

The question that opened this report—"Does it work?"—can be framed more usefully, as the National Research Council committee did in 2004 (Confrey & Stohl, 2004): For whom does it work, under what conditions, and why? A strong impact study answers the first part with a credible comparison and a transparent estimate of the effect's magnitude. The theory of change answers the other two, allowing us to anticipate what will happen when the program is introduced into our classrooms.

The useful question is what needs to happen for this program to work here, with these teachers and students.

In the report "What Impact Do Mathematics Programs Have on Learning?", we apply these questions to the available evidence on mathematics curriculum programs.

Further Reading

The two references marked with ▸ are accessible starting points for readers without a research background who would like to explore further.

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.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Confrey, J., & Stohl, V. (Eds.). (2004). *On evaluating curricular effectiveness: Judging the quality of K–12 mathematics evaluations*. National Research Council, National Academies Press.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350.

Evans, D. K., & Yuan, F. (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis, 44*(3), 532–540.

Every Student Succeeds Act (2015). Pub. L. No. 114-95, 129 Stat. 1802. United States.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505.

▸ Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact evaluation in practice* (2nd ed.). Inter-American Development Bank and World Bank.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

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.

▸ 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.

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), 260–303.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*(6), 304–316.

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

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E., & King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development, 106*, 324–336.

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

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500–506.

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

Williams, R., Citkowicz, M., Miller, D. I., Lindsay, J., & Walters, K. (2022). Heterogeneity in mathematics intervention effects: Evidence from a meta-analysis of 191 randomized experiments. *Journal of Research on Educational Effectiveness, 15*(3), 584–634.

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.

Funding and declaration of interests. *This publication was funded by Innovamat Education and is part of a science communication series on education produced by the organization. Marc Colomer and Carles Aulés are members of the Innovamat team. Caterina Calsamiglia is an external author who contributes her expertise in educational impact evaluation. These affiliations and funding arrangements 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 the people who reviewed this report and helped improve its quality: Ángela Perez, Albert Vilalta, Anna Llobet, Claudia Portas, and Aida Sanuy. We would also like to thank Marta Soler for her work on the layout and graphics.*