DEMANDING MORE FROM EDTECH EVALUATIONS

An Edtech Research Guide for Administrators

RESOURCES FOR VETTING, ADOPTING AND SUPPORTING EDTECH SOLUTIONS

MIND RESEARCH INSTITUTE

Contents

Introduction

In November of 2018, Institute of Education Sciences (IES) Director Mark Schneider issued a message about needing to make research more relevant and usable. His remarks were related to field research sessions conducted by IES and the Jefferson Education Exchange, that were focused on getting feedback from teachers and administrators on topics related to education research. Director Schneider spoke of what he called the "last mile" problem—how to get the information (which is formal, but oftentimes dense or inscrutable) into the hands of the practitioners who need it the most. He stated that his goal is to make information "useful, usable and used."

According to an Education Week article about those recent IES research sessions, there was also feedback about the need for "more webinars, podcasts and similar conversations around research, rather than simply better access to research journals."

At MIND, we completely agree there needs to be more non-academic conversation about research, and that formal research findings can and should be unpacked and translated to become useful to decision makers. But when it comes specifically to research that evaluates edtech effectiveness, the immediate conversation we need to have is less about access, and more about the nature of the information to begin with.

This ebook gathers together content from MIND's podcasts, blogs, webinars, videos and more into a single, easy-to-digest resource aimed at equipping administrators to expect and get more out of current and future edtech research.

For quite some time now, the traditional approach to edtech evaluation—the goals, the conventional process, the vetting, description and use—has been broken. Because of this, administrators can be easily confused or misled by the scarce credible information about program effectiveness. Yet every school year, they still need to make decisions about what tools or programs to source, select, purchase, adopt and support in their schools and districts.

This ebook gathers together content from MIND's podcasts, blogs, webinars, videos and more into a single, easy-to-digest resource aimed at equipping administrators to expect and get more out of current and future edtech research. In the following pages we'll explain what we mean when we say the current system is broken, and provide some clear solutions on how to fix it. We'll break down some key components of research that everyone should have a baseline understanding of. And we'll provide you with resources that will help you ask better questions and make more sense of the edtech research that's out there, including our own.

Why Should You Listen to Us?

MIND Research Institute is a nonprofit social impact organization specializing in neuroscience and education research for over 20 years now. From edtech evaluation to studies on how the brain learns, research is part of what we do every day. Research is literally our middle name, and one of our organizational goals is to make research, generally, more meaningful to educators and administrators. We strive to make all aspects of edtech research more accessible, and better equip educators and administrators to make informed decisions about the varied methods and resources they choose to implement in schools and classrooms.

As part of those efforts, we've created a repeatable, scalable study method that we have been continuously improving over a number of years. We'll dive more deeply into the specifics of this approach in the near future, but the focus of this ebook is on equipping administrators with the knowledge and tools to get more from edtech evaluations right now.

Chapter One:
The Disconnect Between ESSA and Edtech Evaluation

Edtech evaluation is an area of education research that is particularly tricky to navigate, because of a variety of factors. The first is regulatory, and is a case of well-intentioned resources that actually exacerbates an existing problem.

The Every Student Succeeds Act (ESSA) established the Student Support and Academic Enrichment (SSAE) program, the purpose of which was to "improve students' academic achievement by increasing the capacity of states, local educational agencies, schools, and local communities to:

(1) provide all students with access to a well-rounded education;

(2) improve school conditions for student learning; and

(3) **improve the use of technology in order to improve the academic achievement and digital literacy of all students."**

The Problem: The Evidence-Based Checkbox

The ESSA emphasized that interventions being used in schools and districts had to be evidence based, and they established four tiers of evidence (which we'll dive into later). The Department of Education also released some non-regulatory guidance that was intended to help administrators make better-informed decisions about the tools and resources they bring into their schools.

In practice, as long as a product had at least one "good" study associated with it that showed a positive impact on student achievement, a box could be checked and everyone could move on.

But so far this guidance hasn't enabled better decisions because of a persistent second factor that makes edtech effectiveness research so tricky to navigate—the checkbox mentality around it. The guidance published by DOE could have, but did not create a widespread desire to dig deeper into edtech research. Instead it provided a means of

checking boxes in order to justify bringing an edtech program into a school or district. In practice, as long as a product had at least one "good" study associated with it that showed a positive impact on student achievement, a box could be checked and everyone could move on.

Which leads us to a third contributing factor—conventional edtech evaluation research can be overwhelming and time consuming to decipher and digest. Each study is handcrafted by a different team of evaluators for specific research questions, in a specific context, using specific measures. So every one-off study is profoundly, qualitatively, and quantitatively different from the next. Coming up to speed on each new study published is a challenge even for experts, exhausting for laypeople.

In some way, research has to be "standardized" so that after an initial learning curve is climbed, users can easily consume multiple studies.

The Jefferson Education Exchange (JEX) recently conducted a survey of 510 K-12 educators that showed several areas of disconnect between available research and the educator community. According to an article by THE journal on the preliminary results of the study, the majority of educators surveyed accessed research through online searches, rather than sites that collect and curate research. While some educators knew of resources like the National Center of Education Statistics, the Education Resources Information Center (ERIC), or the What Works Clearinghouse, only half had used the ERIC, and less than one-third had ever used the others.

"From our conversations with educators, it seems like they don't have the time to engage deeply, and they need translated and digested research," JEX Director of Implementation Research Emily Barton was quoted as saying in the article.

This barrier to understanding edtech research is a primary reason decision makers take the checkbox approach when evaluating edtech. If we want to change this dynamic, then as IES Director Schneider stated, the research must be useful, usable and used. In some way, research has to be "standardized" so that after an initial learning curve is climbed, users can easily consume multiple studies. We also have to empower decision makers to expect more from edtech research: to not only look to see that research was done at all, but also take a deeper look at what that research is really saying.

In order to do that, we have to first take a look at the current state of edtech evaluation, and where it falls short. With a clearer picture, it's easier to zero in on what you should be looking for.

Chapter Two:
The Problem with the "One Good Study" Paradigm

You probably wouldn't be surprised to hear that every edtech publisher says their product works, and they all have some sort of supporting evidence. But oftentimes that evidence—if it's fully experimental—is very scarce. In many cases, it's just one study. Yet just that one piece of "gold standard" evidence is often considered good enough by educators when making a purchasing decision.

But it shouldn't be.

Educators aren't the only ones stuck in this "one good study" paradigm. Highly credible edtech evaluation lists give top marks for just one RCT (randomized controlled trial). Meanwhile, any other program with studies not meeting the RCT bar of rigor is by comparison downgraded—no matter how many other non-RCT studies they may have, under how many different conditions, in what more recent timeframe, or even with repeated positive results.

The Problem: Gold Standard EdTech Studies are Rare

The biggest problem with relying solely on fully experimental RCT studies in evaluating edtech programs is their rarity. In order to meet the requirements of full experiments, these studies take years of planning and often years of analysis before publication. These delays cause a host of other challenges:

- **The Product Has Changed**: RCT studies can take years to publish, so the product used in the study is often out-of-date. Over time, programs are continually revised to meet new standards, to improve and add product components, or to change (and sometimes reduce) program usage requirements and support. None of these changes are captured by an out-of-date study.

- **All School Districts Are Not The Same**: RCT studies often use only one school district, due to the extensive planning and complexity of a full experimental roll-out and student randomization. But school districts differ in many ways—from technology, to culture, to student subgroup distributions. How do you know the study results are applicable to the specifics at your district?

- **Different States, Different Assessments**: In a single study, typically only one assessment or state test is used, so results are specific to that assessment. But assessments differ widely from state to state in level, type, and emphasis. Results may vary. Moreover, even within one state, the assessment can change from year to year. If the study used a different assessment, how do you gauge the results' validity against your state's current test?

- **Limited Grade Levels**: RCT studies typically cover a specific grade level band, not all grades served by a given product. Yet teachers, usage, and content vary tremendously from K-2, to 3-5, to secondary. Inferring validity of results into unstudied grade bands is questionable due to different levels of content, different teaching methods, different student ages, and different assessments. How do you know the program works for all the grade levels you plan to adopt?

Altogether, these issues demonstrate the often limited relevance of relying on "one good study."

Extrapolating one study's results to your situation is not fully valid unless the study was performed on a district like yours, on the grade band you're planning on using, with a student subgroup mix like yours, with usage like you plan to adopt, on your most recent assessment, and with the program version you'll be using.

School districts differ in many ways—from technology, to culture, to student subgroup distributions. How do you know the study results are applicable to the specifics at your district?

Chapter Three:
Why ESSA's Tier System is Only the First Step

As we stated earlier, the Every Student Succeeds Act (ESSA) provides schools and districts with funding for technology and also includes non-regulatory guidance on how to determine whether an edtech program is effective.

While ESSA's predecessor, No Child Left Behind, also contained a few points around program efficacy evaluations, ESSA takes that one step further and categorizes possible program evidence into four tiers (three for individual studies, and a fourth for other evidence):

Tier	Strength of Evidence	Type of Evidence
1	Strong	Supported by one or more well-designed and well-implemented **experimental studies**
2	Moderate	Supported by one or more well-designed and well-implemented **quasi-experimental studies**
3	Promising	Supported by one or more well-designed and well-implemented **correlational studies** (with statistical controls for selection bias)
4	Demonstrates a Rationale	Based on high-quality **research** findings or positive evaluation that the activity, strategy, or intervention is likely to improve student outcomes or other relevant outcomes

The Problem (Again): One Checkbox Study

But educators shouldn't rely solely on this tier system. An edtech program can be rated as having "strong" evidence based on a single randomized control trial, done many years ago, on a version of a product that no longer exists.

That study might also be on a state assessment that hasn't been around for years, or on a small sample and a specific type of school or student body. And because these ratings do not expire, companies can stop submitting their products for evaluation once they obtain a single "gold standard" study that falls into the first tier.
Because rigorous studies have historically been rare, it's still up to the individual to critically look for a pattern of evidence for each product they are considering.

Aggregated website lists may be a good beginning for someone starting to look for primary source information, but it is still up to the individual educator to ensure that the studies quoted are applicable enough to their school or district, recent enough, and effective based on relevant assessments.

So Where Should Educators Look?

The answer to the question is not so much as a "where to look," but "what to look for." Seeing that a product has one study that falls into the "strong" tier of evidence (or any tier) should prompt you to look deeper. Here at MIND, we believe a high volume of effectiveness studies is the future of a healthy market of product information in education.

Here at MIND, we believe a high volume of effectiveness studies is the future of a healthy market of product information in education.

Chapter Four:
The Lack of Long-Term Accountability

Most educational programs are not being held accountable for long-term impact. This is a problem. The current edtech market is not set up to expect accountability at scale over the long term, but rather to hop from fad to fad. This makes it especially challenging for education leaders to make wise choices, and it inhibits innovation and continuous improvement by program developers.

Let's take a look at three aspects of the problem, which feed into each other:

1. Schoolwide performance data, with its unparalleled ease of access, often can't be used for evaluations because of haphazard use commitments. Once an innovative new program is chosen, it is often not used broadly enough across an entire school grade level, nor at scale across a whole district. This makes a statistical analysis based on those easily accessed state-published schoolwide scores impossible, leaving only student-level studies which, due to data sharing protections, are rare, expensive, and slow to publish.
2. When the district and school aren't fully committed to a program adoption, teachers and staff don't receive the time and support they need to consistently use the new program with fidelity. Failure to meet minimum program implementation requirements, including sporadic use, means program outcomes will not be achieved.
3. New program adoptions that lack proper upfront training don't get widespread use, aren't used enough, and can't show full results, so they are often abandoned after a few years. This prevents multi-year, longitudinal program evaluations. No generalizable findings, not even about success factors for adoption, are gained. Program selection can't become a continuously improved core competence. Program selection is driven by who's selecting, rather than about making informed decisions. All that's left is unfortunately very conventional and common: chasing the newest, most hyped fad.

When the district and school aren't fully committed to a program adoption, teachers and staff don't receive the time and support they need to consistently use the new program with fidelity.

All of the above conspire to inhibit continuous improvement. Currently, a multitude of educational programs' claims about usability, effectiveness or even their pedagogical features are very similar, very loud, and very confusing. They lack a basis for validation. They are here today, gone tomorrow. The market is in dire need of a better path for:

- Districts to select, pilot, implement and evaluate programs; and
- Publishers to use evaluations to improve fundamental usability and effectiveness at scale for all student subgroups.

At MIND Research Institute, we have been leveraging the efficacy advantage of our unique visual product design, our long-term district adoption track record, and our growing list of blue-chip research partner organizations, to pilot another way. We believe educational leaders can always be provided sufficient evidence to reasonably and robustly evaluate and predict program usage and performance in districts like theirs at scale, for their specific mix of students and teachers.

So what does it look like if a product provides schools and districts with repeatable, predictable, long-term impact for all groups of students? This is a question our school partners, teachers and students have been answering at scale over the last ten years.

It is the collective responsibility of all educational program providers to share their program results with authenticity, transparency and clarity. This is our leadership model in providing educators with what they need to evaluate and choose effective programs.

We believe educational leaders can always be provided sufficient evidence to reasonably and robustly evaluate and predict program usage and performance in districts like theirs at scale, for their specific mix of students and teachers.

Building for the Long Term

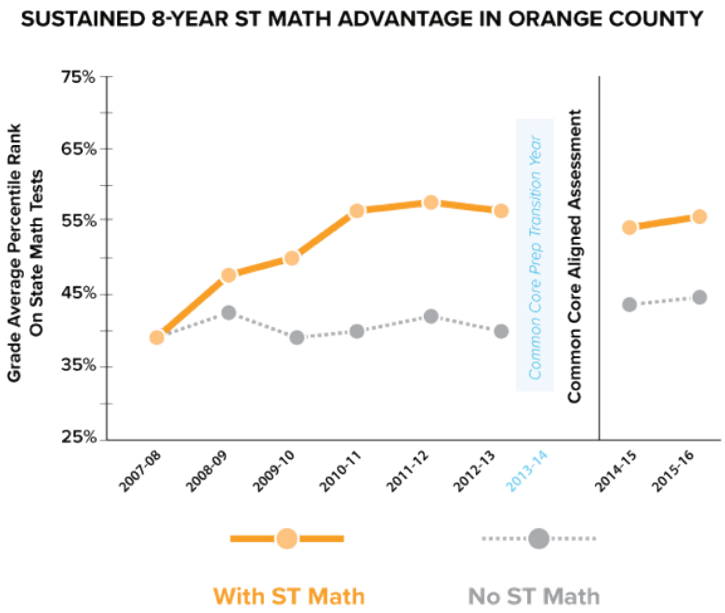
The partnership between ST Math and Orange County, California schools is a proof of concept that shows the impact an effective program can have on a diverse county of 3.5 million. We are proud of the results our great district partners have achieved in fidelity of program use at scale, in engaging diverse student subgroups and in state test score advantages, evaluated consistently over a ten-year period.

But it didn't start out that way. Initially in the OC area, only a handful of new schools were adopting ST Math each year. Fortunately, those adoptions achieved grade-wide use, the teachers loved ST Math and the program was implemented with fidelity. But rapid growth wasn't happening. Even though the early adopters achieved great results on schoolwide tests, adjacent principals typically attributed that success to unique leadership or a high performance staff that couldn't be replicated elsewhere.

Crucially, there weren't enough schools using ST Math to generate credible, rigorous statistical studies—to provide the evidence of how the program could impact all teachers and students in the county.

As a non-profit, we decided to approach local business and community partners to help proactively spread this program to schools who needed help with math. These visionary philanthropists, intrigued by the early results, said yes. Thus in Orange County began our ST Math School Grants program, which continues nationwide to this day.

Our partners eventually contributed the grant support to spread ST Math to over 100 local low-performing schools within a two-year span. Because of the results they saw in their classrooms, these districts opted with their own budgets to continue the program after the end of the first grants-supported year. Since then, over 95% of grantee schools have continued to use the program.



The chart above shows an 8-year long analysis. In 2007/2008, the ST Math schools were matched to similar schools. Over the next 8 years, ST Math schools opened up a 10-15 point consistent advantage in statewide rank.

Demand Long-Term Accountability

As administrators are evaluating potential programs and resources to bring into their school or district, they should be thinking of the past, present, and future. Look for and ask for impact studies that demonstrate long-term efficacy of a program. Are there sustained results in other schools or districts like yours?

You also want to inquire about the methods and frequency for evaluating program impact. How will results be measured consistently, over a period of time?

We've continued to evaluate the impact of ST Math over that time as well. You can see the latest results over at stmath.com/impact/results.

Chapter Five:
Repeatable Results at Scale

The time has come for a shift in how we evaluate edtech programs. Rather than relying on one "gold standard" study, we should be looking at a large number of studies, using recent program versions, garnering repeatable results, over many varied districts. Quasi-experiments can study the adoption of a program as is, without requiring the complexity and time that formal experiment planning takes. Methods of matching and comparing similar schools with and without the program can be made statistically rigorous and powerful. And if we study at the grade-level, we have the test performance data universally available on state websites. It is then possible to do a quasi-experimental study on any large enough school cohort. If, instead of relying on RCT's alone, we pay attention to quasi-experiments, a much higher number of studies is possible.

Crucially, a larger number of studies enables buyers to evaluate repeatability. Why is repeatability so important? Because even the "gold standard" results of a single study in the social sciences have very often failed to be replicable. In fact, published studies are allowed a 5% chance of drawing a false conclusion; how does one know that one "gold standard" study was not itself a cherry-picked or fluke result? So, "one good study" is not enough evidence. Reliable results as evidenced by replication from a lot of studies need to be the new normal.

If, instead of relying on RCT's alone, we pay attention to quasi-experiments, a much higher number of studies is possible.

Imagine this new paradigm with a large number of recent studies—let's say five or more. This can allow us to look for consistent patterns over multiple years, across grade levels, and especially across different types of districts and assessments.
This paradigm shows its rigor through repeatability, and adds vastly improved validity with respect to:

- Recent version of the program, training and support;
- A real-world variety of types of use, districts, grade-levels, teachers, and student subgroups;
- Patterns of results across many different assessments.

At MIND, we believe a high volume of effectiveness studies is the future of a healthy market of product information in education. To illustrate and promote this new paradigm, we've created a program evaluation rubric.

Evaluative Element	Low	Medium	High
Different Assessment Instruments Used in Studies	Only 1	2 - 3	4 or more
Sample Size	< 100 students, 1 school	< 500 students, 1-5 schools	6 or more schools
Repeatability	< 3 studies	3 - 5 studies	6 or more studies
Applicability to Student Subgroups	No results broken out by student subgroup	Results reported for some subgroups of interest, but not all	Results evaluated for all major subgroups of interest for intended use
Variety of Schools	All schools evaluated are similar to one another (size, locale, district, demographics)	Studies evaluate different types of schools	Results from multiple studies cover school types similar to intended use
Range of Grades	Studied grade range is different from intended use	Studied grade range is similar to intended use	Results cover all grades intended for use
Study Controls	No control results are reported; only treated school results	Control results are referenced, but were not rigorously matched	Rigorously matched control results reported
Independence	No independent third party studies	1 - 2 independent third party studies	3 or more independent third party studies
Product Relevance	Study covers an old program version that is substantially different from current program	Studies are 3 or more years old but cover largely similar program revisions	Studies are updated every school year and cover the current program revision

= MIND Research Institute

While MIND has not yet achieved the highest standard in each of these rubric sections, we are driving toward that goal as well as annual, transparent evaluations of results of all school cohorts. We've already been able to do just that in grades 3, 4 and 5. We want our program to be held accountable for scalable, repeatable, robust results—it's how the program will improve and student results will grow.

Effective learning is important enough that there should be studies published every year covering every customer.

Chapter Six: Takeaways

Remember, most people who are buying a new car are not mechanics or automotive experts, but there are certain things people generally know to look for from their time of automobile experience, things beyond showroom shine and price. We know to inquire about variables like gas mileage, number of seats and space, power, comfort, safety rating, reliability and what kind of warranty is offered. There is a baseline level of knowledge needed to make such a major financial decision. Otherwise, we risk being saddled with what was once a shiny new car, but is now an expensive mistake that doesn't meet our needs.

Similarly, in order to make well-informed decisions about edtech solutions for your school or district, you have to inquire about features below the surface gloss. And unlike the mature and standardized automotive measures like EPA mileage numbers, or NHTSE crash safety ratings, edtech measures are immature. The fundamental offer of edtech cannot be taken for granted: will your people be able to use it effectively, and under what conditions will it reliably get all of your students to a desired destination. While you don't have to be an expert on research and evaluation, you do need to know what kinds of things to look for, and you do need to ask for them to be explained in layman's terms.

While you do not have to be an expert on research and evaluation, you do need to know what kinds of things to look for, and you do need to ask for them to be explained in layman's terms.

Questions to Ask Up Front:

How many students of what type does this study involve?

If the study is on a small group of students, and it's the only study available, that should be a red flag. For our own evaluations of ST Math, we've done a high volume of studies with student groups of different sizes, some of which include tens of thousands.

To get specified study results what are the minimum usage requirements?

For ST Math, we require a consistent investment of 90 minutes per week in the program, which results in covering at least 50% of on grade-level content.

What assessments are used?

Is the study based on a proprietary assessment tool created by the program itself? A third-party assessment, but not a high stakes state test? For ST Math, we have standardized on use of publicly available assessments of schoolwide math performance.

What version of the program is the study on?

If a study is older, it may be on a version of the product that doesn't even exist anymore. Look for recent studies on the current version of the program.

Is it a matched comparison study?

To infer cause and effect, a study must compare two groups' performance, where the difference is the program use.

Interpreting Results

As we noted before, every edtech product has a study that says it's effective. But what are the impact results of that study really saying? How meaningful is the difference between outcomes? And how do we know that difference is not due to chance? The effect size and statistical significance of the findings can help you answer those questions.

Effect Size

Simply put, effect size is a way of describing in a standard way the magnitude of the difference between two groups. Effect size calculations generate a "difference" figure of merit that takes into account "variance." Effect size is what statisticians use as the benchmark of how much difference there is between two groups. It lends itself to becoming a standard scale to judge "size" of any program's impact. It's a way to use the same measuring stick across any study

for evaluating the importance of a difference between any two groups – even if the study measures are very different. Crucially, effect size takes into account sample variance as well as the difference in the averages of the two groups. Effect size is calculated from the sample variance, and the difference in the averages of the two groups.

For more resources on effect size, you can watch our video "What Is Effect Size?" and listen to our "Understanding Effect Size" podcast.

Understanding Statistical Significance

Statistical significance is about the probability that some observed difference between two groups could be due to chance. This is a special application of the work "significance," we sure wish statisticians had chosen a more unusual word! Because in this application, it doesn't mean "important," it just means that we're pretty sure (95% usually) that we weren't fooled into thinking there was a difference, when there really wasn't. When it comes to statistical significance, since it's the chance we have been "fooled" by the difference, the lower the number is, the better. So, if you have a statistical significance score of .02, that means there is just a 2% chance the difference is due to chance. This statistical significance metric is also called the "p-value," and is calculated using both sample size, sample standard deviation, and difference in sample means.

In Conclusion

As you navigate the roiling, muddy waters of edtech evaluation, the number one thing to remember is that you should expect more. One study is not enough, one subgroup is not enough, even one assessment type is not enough. If a program is truly effective, you should expect to see the following five things:

1. A large number of studies
2. Study and results replicability
3. Studies on recent versions of the program
4. Effect sizes as reflected by multiple state assessments
5. Studies with many varied districts and school situations, subgroups and usage models

In order to know whether a program will be effective for your district, your schools and your students, you need to look for studies (multiple studies!) that reflect your population, your assessments, and the version of the program you are looking to start using.

We believe MIND is ahead of the curve on study volume, but we are not alone. For example, districts are starting to share the results of their own many studies with peers on LearnPlatform, which will eventually create a pool of studies covering a variety of conditions for any program. With more studies, more replications, and more varieties, there will be more relevant information to inform purchasing decisions, and educators will find better product fit for their teachers and students.

One study is not enough, one subgroup is not enough, one assessment is not enough.

If you would like to see more about our results, visit stmath.com/impact/results. If you'd like to hear more about edtech evaluation, we have the following podcast episodes that dive into several of the topics covered in this ebook:

A New Paradigm for Edtech Evaluation
Repeatable Results
Effect Size

As mentioned in the introduction, in future conversations we'll be diving deeper into our repeatable and scalable approach that can help systematize edtech evaluations, so rather than peering into muddy waters, administrators will be able to see everything clearly.

More from MIND

At MIND Research Institute, our mission is to ensure that all students are mathematically equipped to solve the world's most challenging problems. So, when implementing our approach to learning, we focus on creating mathematical content to serve that mission.

In the classroom, our PreK-8 visual instructional program ST Math focuses on dynamically activating our students as learners. ST Math builds a deep conceptual understanding of math through rigorous learning and creative problem solving.

To learn more about ST Math, visit us at stmath.com or click the link below to request information on the program.

LEARN MORE ABOUT ST MATH

ST Math

You can also see and hear more from us on the MIND blog (blog.mindresearch.org), which features podcasts, thought leadership articles, interviews, resources and more.

About the Author:

Andrew Coulson
Chief Data Science Officer

Andrew Coulson leads a team of data analysts to conduct evaluations of student and teacher usage and outcomes of ST Math. He previously led MIND's education division for 12 years, where he helped to execute strategies to scale the organization's reach to over one million students. Prior to joining MIND, Coulson was the first education program officer for a major Orange County foundation, and also worked for 17 years in upper management in high-tech manufacturing engineering. Coulson holds a master's degree in physics from UCLA.

Editor: Brian LeTendre
Designer: Jo Zafra

MIND RESEARCH INSTITUTE