MIND RESEARCH INSTITUTE

Public Comments on Proposed Priorities
for the Institute of Education Sciences,
U.S. Department of Education | May 2019

Thank you for the opportunity to provide public comments on proposed priorities for the U.S. Department of Education Institute of Education Sciences (IES) [84 FR 11755].[1]

MIND Research Institute[2] is delighted to have the opportunity to share our perspectives as a national non-profit organization that is focused upon ensuring that all children are mathematically equipped to solve the world's most challenging problems.

Who We Are

MIND Research Institute is a nonprofit, neuroscience, social impact organization. We are innovators committed to transforming education and closing the experience gap for all learners to raise the next generation of critical thinkers and problem solvers.

MIND was founded in 1998 by three University of California researchers united behind a simple yet innovative idea: let's teach math the way children learn—visually and experientially.

ST Math[3] – our PreK-8 visual instructional program – leverages the brain's innate spatial-temporal reasoning ability, allowing students to begin solving mathematical problems visually and incorporating symbols and discourse as their understanding deepens. Because of this approach, the program is highly effective for all children, regardless of socioeconomic, linguistic, or cultural background.

Following, we present our comments on a subset of the goals outlined in the IES *Federal Register* notice dated March 28, 2019.

[1] https://www.federalregister.gov/a/2019-05970

[2] https://www.mindresearch.org/

[3] https://www.stmath.com/

IES Proposed Goal:

To better measure and understand the variation in the effectiveness of education programs, practices, and policies.

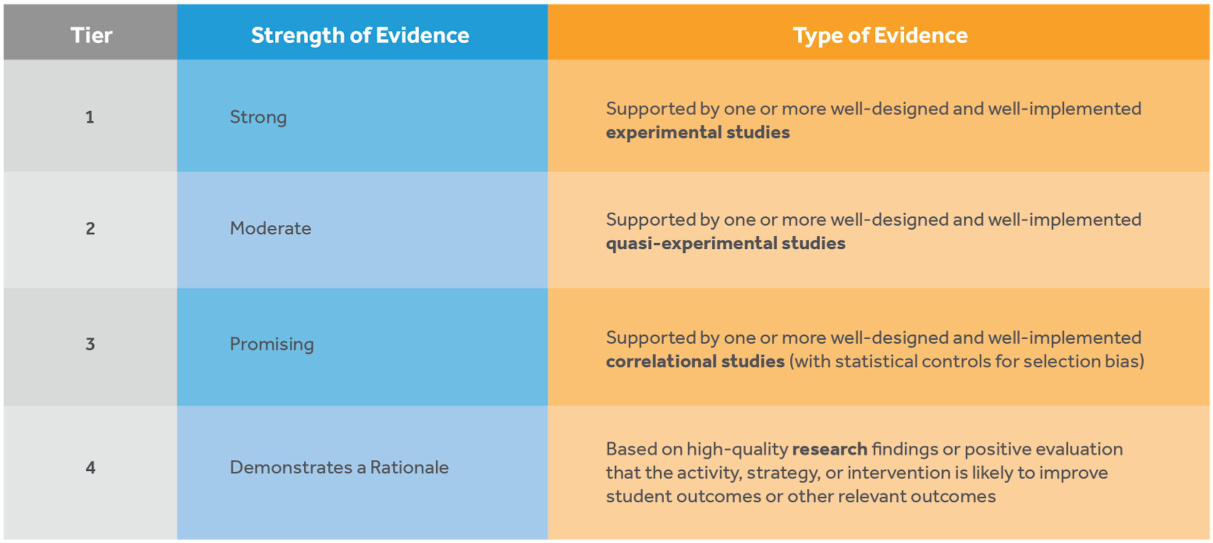
MIND commends IES in setting this goal, which represents a crucial step forward in improving the health of the edtech evaluation market.

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 guidance and resources that we believe have actually exacerbated an existing problem, which is the tendency to fall back on what we'll describe in this comment as a single study "checkbox" paradigm.

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."

ESSA emphasized that interventions being used in schools and districts had to be evidence-based, and they established four tiers of 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

DOE also released some non-regulatory guidance[4] that was intended to help administrators make better-informed decisions about the tools and resources they bring into their schools.

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 original guidance published by DOE could have, but did not create a widespread desire by educators to dig deeper into edtech research. Instead it provided a means of merely checking a box in order to justify bringing an edtech program into a school or district. In practice, as long as a product had at least one rigorous study associated with it that showed at least some positive impact on student achievement, educators could feel they had checked the box and everyone could move on. The new goal articulated here by IES is positive step toward breaking this checkbox paradigm.

What's Wrong with Checking the Box?

Put simply, we believe that educators shouldn't rely solely on the ESSA tier system until it also includes an expectation and then the reality for *multiple* studies for each product, at tiers 1-3.

As it currently stands, 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 so rare, it's still up to the individual to critically look for validity of the evidence for each product they are considering. This new IES goal is a step towards making it easier for educators to find that pattern.

As written, the goal implies support for many studies, in stark contrast to the current paradigm of a scant one or two studies per program spread over a decade or so. That paradigm, which is in part caused by a scarcity of studies, presents a number of challenges, including lack of variation in subjects, implementation variables, and assessments, as well as a lack of replicability evidence, due to most studies not being replicated.

At the level of theory, this goal rightly implies that different studies will identify different effect sizes - of course! So, there is not one "right" number, and some spread in effectiveness outcomes is to be expected, and needs to be evaluated.

[4] https://www2.ed.gov/policy/elsec/leg/essa/index.html

Further, consider that different studies may identify a wide range of effects and will also have different p-values. The old paradigm of yes/no based on p-value of one study should also be replaced with more nuanced compatibility intervals.[5]

Making It Easier for Educators

Aggregated website lists of ESSA tiers, or the What Works Clearinghouse (WWC), 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, beyond the word "strong" and beyond a "green" icon, the studies quoted are applicable enough to their school or district conditions, are recent enough, and are effective based on relevant assessments.

The solution for educator is not so much about "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 nonetheless prompt them to look deeper and discover more.

> ***IES Proposed Goal:***
> *To disseminate the results of scientifically valid research, statistics, and evaluations in ways that are accessible, understandable, and usable in the improvement of educational practice by teachers and other educators, parents and families, learners, administrators, researchers, policymakers, and the public.*

It's certainly true that conventional edtech evaluation research can be overwhelming and time-consuming to decipher. Each study is hand-crafted by a different team of evaluators for specific research questions, within 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 for experts, and is exhausting for lay people.

The Jefferson Education Exchange[6] (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[7] on the 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.

[5] See "Scientists rise up against statistical significance" at https://www.nature.com/articles/d41586-019-00857-9. Co-signed by nearly 800 signatories, the article calls for a stop to the use of p-values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis.

[6] http://jexuva.org/

[7] https://thejournal.com/articles/2018/11/27/how-educators-utilize-research.aspx

"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 told THE Journal in another recent article about the survey.

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. MIND Research Institute believes that, in some way, **the research has to be "standardized"** so that after an initial learning curve is climbed, users can easily consume multiple studies - and the WWC listings could be part of the standardization effort, as described in the next section of this comment. 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 take a deeper look at what the specifics of the study and the results are really saying.

> ***IES Proposed Goals:***
> *Enhance the experience of What Works Clearinghouse users, adding features that make its reviews more useful and usable.*
>
> *Increase the number of What Works Clearinghouse Practice Guides and Intervention Reports, ensuring that they are written in an accessible manner and supported by material that increases the use of this information.*

The What Works Clearinghouse has been a crucial precedent for compiling the results of rigorous evidence, to be sure. When thinking about ways it can be improved, it may be helpful to think of an analogy to the car-buying process.

Most people that are buying a new car are not mechanics or automotive experts, but there are certain things people generally know from a lifetime of automobile experience to look for beyond showroom shine and price when evaluating different cars for purchase. We know to inquire about independent 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 we need to take into these major financial decisions, or we risk buying and being saddled for years with a pretty car that is not going to 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 in the mature and standardized automotive measures like EPA mileage numbers, or NHTSE crash safety ratings, edtech remains an immature market. The fundamental offer of any edtech solution - 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 - cannot be taken for granted. 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.

But the pendulum shouldn’t swing too far in either direction. In its attempt to be quick and simple, we feel that the WWC format falls short on credibility, validity, and nuanced understanding. Reliance upon a term like “strong” or a symbol like a green traffic light can make it difficult for educators to understand and compare efficacy, and make an accurate determination about whether the edtech solution will work for their students. We recommend that WWC reporting be very explicit about the following factors - in a standardized way - that will help educators to conduct their own analysis:

- *What is the type of study?*

- *What are the student demographics for this study, and what type of district or school is involved?*

- *Does the study offer specific outcome metrics and their range of effect sizes, not just one average?*

- *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 caution 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?*
  Program implementation information (including support) and dose thresholds are key data points for educators to review. For example, for ST Math, our protocol requires a consistent investment of 75 minutes per week in the program, which results in students on average covering at least 50% of on grade-level content.

- *What assessments are used?*
  Is the study only involving an assessment tool created by the program itself? By a third party? For ST Math evaluations, we have standardized on publicly available state assessments of schoolwide math performance.

- *How old is the study, and 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. Or it may be an out-of-date implementation model or support process. And beyond the product version itself, have there been changes since the study in standards or assessments? Look for recent studies on the current version of the program.

Interpreting Results

As we noted before, every edtech product has a study that says it’s effective. But what are the results of that study really saying when you take a deeper look? How meaningful is the difference between outcomes for the group of students who did not use the tool, product or program, and the one that did? What exactly did the “controls” do instead of the treatment? For that matter, what other things did the treatment group do that also affects scores? And how do we know that difference is not due to chance? The effect size and statistical significance of the findings can help educators answer those questions.

Explaining these concepts in layman's terms[8] will go a long way in **helping administrators to use the data to make the best possible decisions** for their schools and districts.

IES Proposed Goals:

Develop and refine education research methods including new methods that take advantage of large administrative data sets and increased computing power.

Expand the use of research using longitudinal data sets.

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.[9]

Quasi-experiments can study the adoption of a program as is, without requiring the complexity and time that up-front 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 large administrative data sets mentioned: the average 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 implying that only costly, lengthy, rare randomized control trials (RCTs) are quality evidence, **we encourage and value tier 2 quasi-experiments**, then 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 RCT study in the social sciences have very often failed to be replicable. Certainly, the measured effect found in just one study is unlikely to be the "true effect," and numerous replications are required to begin to form a conclusion about the true effect. Moreover with so few studies published, how does one know that one "gold standard" study was not itself a cherry-picked 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.

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:

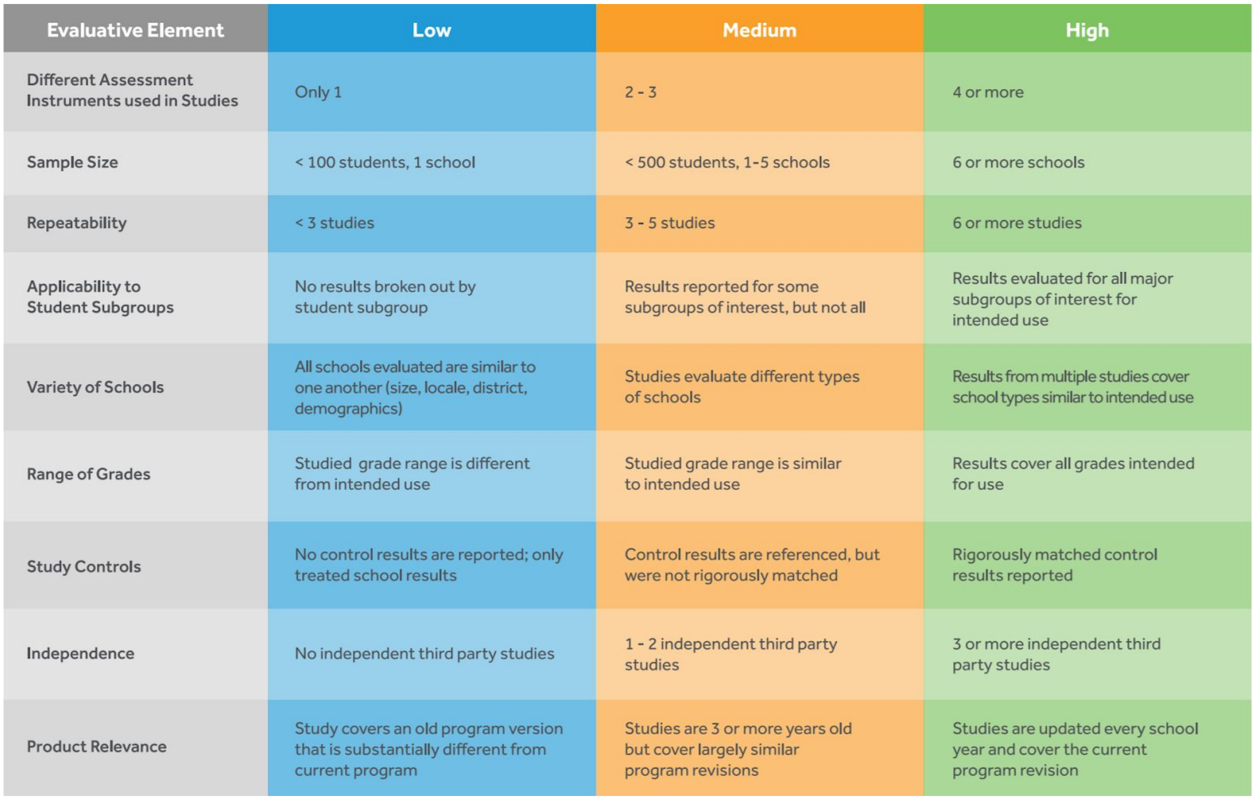
- Recent version of the program, training and support;

[8] What is Effect Size? https://blog.mindresearch.org/blog/what-is-effect-size

[9] See "Debunking the 'Gold Standard' Myths in Edtech Efficacy," EdSurge. https://www.edsurge.com/news/2019-05-21-debunking-the-gold-standard-myths-in-edtech-efficacy

- 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

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 ST Math 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 school.

That's why, in addition to third-party validation, the MIND Data and Evaluation team performs multiple annual, transparent evaluations of results of all ST Math school cohorts.

IES Proposed Goal:

Encourage partnerships between researchers and private companies, both non-profit and for-profit, to put interventions that work into more schools and in the hands of more teachers, parents and families, and learners.

We strongly support IES' plans to encourage partnerships that will facilitate the broad-based implementation of edtech programs for applications that are proven to work under specified conditions for students.

An initial observation is that the incentives and motivations of academic researchers can be on a different plane from the needs of the market. The phrase "interventions that work" should be unpacked and better defined, so it's clear what IES is encouraging. For example:

- Interventions that work how well?
- Under what conditions? When used how?
- For which students?
- At what cost in time and money?
- Over what time period?

And in the cases where more than one product is deemed to "work," then what is the department's guidance for school administrators? Does this return to the notion of just picking the highest single-study effect size? We believe strongly that IES should facilitate the greatest volume of effectiveness studies and data that are presented in a consistent and easy-to-understand format - all with the goal of helping administrators to make the best possible decisions for their schools and districts.

Thank You

On behalf of MIND Research Institute, we thank you for your consideration of these comments. Requests for clarification or additional information may be directed to Liz Neiman, vice president of engagement, at eneiman@mindresearch.org.