Guidelines for Conducting and Reporting EdTech Impact Research in U.S. K-12 Schools

Denis Newman, Andrew P. Jaciw, and Valeriy Lazarev

Empirical Education Inc.

April 15, 2018

Empirical Education
EMPOWERING EDUCATORS THROUGH EVIDENCE AND INSIGHT

ETIN
Education Technology
Industry Network of SIIA

About the Education Technology Industry Network of SIIA

ETIN is the leading voice for companies that provide software applications, digital content, online learning services and related technologies across the PK-20 sector. ETIN drives growth and innovation within the industry by providing leadership, advocacy, business development opportunities, government relations, and critical edtech market information. SIIA is an umbrella association representing 800+ technology, data, and media companies globally. For more information, visit siia.net/etin

About the Authors

Denis Newman is the CEO and founder of Empirical Education Inc., which has pursued a mission of helping educators make evidence-based decisions. Newman has 35 years of experience studying student-teacher learning processes and developing instructional technologies. His Ph.D. is in Developmental Psychology from the City University of New York. He has conducted research and development at Rockefeller University, UC San Diego, Bank Street College of Education, and BBN Corporation. As a pioneer in the application of Internet technologies for student learning, professional development, and school administration, he is widely published and has served as program chair for the American Educational Research Association's Curriculum and Learning Division. His business career has included senior positions at educational software companies Tegrity and Soliloquy Learning.

Andrew Jaciw has 20 years of experience in the field of education: 6 as a practitioner and 14 as a researcher. As Empirical's Chief Scientist, Jaciw has been the primary developer of the experimental and analytic designs for the company's experimental research. For evaluation of school programs where implementation issues are critical, he applies moderator and mediation analysis and approaches to estimating impact on those receiving treatment with sufficient dosage. Most recently his research has focused on empirical study of the relevance and reach of causal impact findings from experiments. Jaciw leads several i3 evaluation teams. He has extensive experience using hierarchical generalized linear models applying SAS, R, and HLM software. Stanford University awarded Jaciw an MS in Epidemiology and a PhD in Education. Before his MS, Jaciw earned a BS in statistics and MA in math education at the University of Toronto.

Valeriy Lazarev holds a PhD in economics and has held research and teaching positions at Stanford University, Yale University, and the University of Houston. His expertise includes econometrics, educational effectiveness and policy analysis, and quantitative study design. Dr. Lazarev has been a lead investigator in many projects funded by IES, the Gates foundation, and other agencies, including studies on teacher effectiveness commissioned by Regional Education Labs. His most recent work focuses on Empirical's research services for edtech companies. He is an active member of American Education Research Association (AERA), Association for Education Finance and Policy (AEFP), and Society for Educational Effectiveness (SREE), and has extensively presented his research in the conferences of these groups.

About Empirical Education Inc.

Empirical Education Inc. is a Silicon Valley-based research company that develops tools and services to provide the evidence K-12 school systems need to make evidence-based decisions about their programs, policies, and personnel. The company brings its research, data analysis, engineering, and project management expertise to customers including edtech companies and their investors, the U.S. Department of Education, foundations, leading research organizations, and state and local education agencies. Over the last decade, Empirical has worked with school systems to conduct dozens of rigorous experiments and now offers services to edtech companies for fast turn-around and low-cost impact studies of their products. For more information visit https://www.empiricaleducation.com.

Acknowledgements

This new edition of the research guidelines is a revision of the guidelines originally published by SIIA in 2011, which was a collaboration of Dr. Newman and ETIN's Research & Evaluation working group, co-chaired by Rob Foshay (Texas Instruments) and brought together by Mark Schneiderman, then SIIA's Senior Director of Education Policy.

For their thoughtful comments on earlier drafts of this current edition, SIIA and the authors thank the following people without in any way blaming them for errors that remain:

- Karen Billings, BillingsConnects
- Mahnaz R. Charania, Independent Consultant
- Myron Cizdyn, The BLPS Group
- Andrew Coulson, MIND Research Institute
- Christine Fox, SETDA
- Christine Gouveia, McGraw-Hill Education
- Amar Kumar, Pearson PLC
- Christina Luke, Digital Promise
- Alexandra Resch, Mathematica Policy Research
- John Richards, Consulting Services for Education, Inc.
- Anne Schreiber, Renaissance Learning, Inc.
- Robert E. Slavin, Johns Hopkins University
- Anne Wujick, MDR

We also thank Robin Means, Kylene Shen, and Hannah D'Apice for their astute editorial advice.

Table of Contents

INTRODUCTION

SIIA originally published guidelines for research in 2011. This substantial revision was published in the summer of 2017 with a commitment to continuous updating to keep pace with changes in government regulations, research methods, and the kinds of data available for research. The current version was released in spring 2018. The original set of guidelines was created at a time when the No Child Left Behind (NCLB) Act had made "scientifically-based research" a buzz-word and the randomized control trial had been introduced to education as the "gold standard" for research. Three major changes have occurred since then that drive the need for these updates.

1) The pace of development and product release has accelerated, reducing the shelf life of research reports and putting greater emphasis on using information from the field in the development and revision cycles.

2) The movement of new software to the cloud is providing much greater access to usage data that gives researchers a clear definition of who the users are and their extent of use.

3) The passage of the Every Student Succeeds Act (ESSA), which includes a clear definition of what counts as evidence of impact, has expanded the definition of useful research.

Many other changes have taken place in the last half dozen years. These include the proliferation of increasingly powerful portable electronic devices, declining demand for conventional print materials, and significant recovery from the economic downturn.

These Guidelines focus on the K-12 market rather than higher education or pre-K centers because there are characteristics that make schools unique, particularly the availability of standardized data at national, state, and school district levels that can be used as outcome measures. These characteristics are driven by federal legislation in the U.S. that does not regulate pre-K, higher education, or international institutions.

We believe the one thing that continues to grow is interest—at state and local levels—in evidence to support procurement of edtech products (in this document we will use the term "product" broadly for any kind of software or network-based instructional or infrastructure product or service provided to K-12 students, teachers, schools, or education agencies). We see the influence of ESSA in the evidence standards being included in state plans (Results4America, 2018). The federally-funded Knowledge Utilization Centers have documented the extent of research being used in district and building-level purchasing decisions (Penuel et al., 2017). The need for evidence is being mentioned by school administrators who are overwhelmed by the number of software applications available to perform school functions. Questions being asked include: (1) “Can it work in schools like mine?”; (2) “Does it qualify for the grant program for which I’m applying?”; (3) “Will it improve our rate of disciplinary referrals?”; (4) “Will I see evidence of student growth?” The growth rate of this interest is perhaps matched by the slow growth of useful evidence to support procurement of edtech products, a rate that we hope to accelerate through these Guidelines.

Evidence of the effect (or potential effect) of products on specific goals—in the specific context of a school system’s population, resources, and challenges—is the kind of information that can be determined by studies of impact. There are many other kinds of questions that educators ask about ease of implementation, bandwidth requirements, alignment with standards, and teacher acceptance, all of which are important in decisions leading to the acquisition of an education technology product. These Guidelines are focused on the ways to show impact on the desired or promised outcomes. They include discussion of not just what has been considered the scientific “proof” of a causal connection between the product and the outcomes, but also ways that research can show that a product has promise or a promising rationale; that is, worth trying out since it has characteristics that are likely to have an impact.

Research on edtech products shares many features with research on other domains such as workforce development, social policy, and medicine, but it works within a unique context. To provide information of greatest value, the Guidelines focus on issues that are pertinent to edtech products and may not be particularly well addressed in conventional literature on evaluation design and methodology.

After an introduction to the purpose and current context of the edtech marketplace, we present 16 guidelines in four clusters representing the phases of a research study, from “Getting Started” through “Designing the Research” and “Implementing the Design” to “Reporting the Results.” These guidelines are not written as a definitive set of standards with a level of specificity to which research reports should adhere; however, they are meant to serve more as a comprehensive reference for what should be included in the planning, design, implementation, and reporting of research. There is a mixture of recommendations for providers and suggestions for researchers, but in all cases, the intent is to inform providers of what they need to know to successfully navigate the market’s need for evidence.

Purpose and Audience

The Guidelines have several purposes and audiences within the K-12 market.

Edtech Providers

The Guidelines are intended primarily for developers, publishers, and service providers of K-12 edtech products (we will use the term “providers” throughout this document). Many edtech providers have already responded to the need for evidence by enhancing the scale, scope, and rigor of their existing research investments, including further documenting the research basis of their products and offerings, and commissioning additional evaluation research. While some providers employ experienced researchers, many companies do not have in-house expertise. The target audience for this document is the subset of company decision makers who are concerned with evidence, but are without research expertise—that is, the managers responsible for development, evaluation, and marketing of these products. Thus, the Guidelines address operational decisions—planning, designing, conducting, and reporting—that are under the control of providers carrying out or commissioning a study. These are practical recommendations to help plan for evaluations, illustrate best practices, and assist in evaluating the potential offerings of outside researchers who might be contracted to conduct the studies.

Educators

We hope that the Guidelines will give K-12 educators confidence that providers understand the importance of presenting information that is unbiased, actionable, and of the greatest value in helping them select and implement edtech products. Research reports that adhere to these Guidelines can be expected to be of high quality and credibility and, therefore, the reports of the evaluations will have utility to education decision makers.

The Research Community

Finally, we hope that the Guidelines will be reviewed by additional stakeholders: researchers, policy makers, and education officials. For all audiences, the Guidelines not only provide approaches to practice, but also seek to advance the field by helping to identify an appropriate balance among the rigor, practicality, timeliness, and usefulness of evaluation studies of K-12 edtech products.

What's New Since 2011

We start these Guidelines with background considerations and observations of the current context of edtech development and deployment. This includes: a) consideration of the pace of change and need for an approach to research that can keep up with it, b) issues related to cloud-based data, and c) an outline of what ESSA says about evidence, which provides a useful framework for the recommendations.

Accelerating Pace

The pace with which technology products are changing and improving has accelerated since 2011. Moving edtech applications online has made it possible for providers to rapidly distribute products to large numbers of customers. With states moving assessments online, schools have hastened their acquisition of bandwidth and one-to-one programs. The 5,000 edtech products reported available for K-12 schools have increased the need for online infrastructure tools and have helped to make processes—from IT to purchasing—more efficient.

This continued rapid development cycle is putting demands on research that are only beginning to be addressed. With conventional evaluation studies, several years may pass between the initial stage of identifying participants and the final stage of reporting results. By the time research is completed and distributed, the technology products may no longer be available in the format or version studied. By contrast, today, the shelf life of a research report may be no more than a couple of years. This situation has two important consequences that are explored and partially addressed in these guidelines. First, a research study is as likely to be useful for improving the product as it is for measuring the product's impact; the study may contribute to both. Second, the "holy grail" of getting research published in a peer-reviewed scientific journal—while still important for scholarly and scientific contributions, and as background for establishing the rationale for a product—is no longer appropriate for reporting evidence about edtech products, where shelf life is shorter than the multi-year review process. We also see growing interest in what is called "rapid cycle evaluation," by which schools pilot (see Glossary) products then conduct studies themselves, eliminating the bureaucratic overhead and custom planning that conventional research usually requires. The same approaches have been used in provider-initiated research, such as Empirical Education's Evidence as a Service.

Move to the Cloud

The last decade has seen an acceleration in the movement of content and computing to internet servers, popularly referred to as "the cloud." Portable devices—such as the Chromebook, iPad, and other tablets—designed for cloud computing have set new standards for instructional delivery by supporting personalized learning (through individual logins and tracking), and facilitating frequent software and content updates that accelerate development cycles. The opportunities and issues raised by cloud computing were not addressed in the original Guidelines but now play a major role in impact research (see Glossary), including tracking product implementation, and analyses of what's working for whom, and under what implementation conditions. While usage and learning-process data provided by cloud-based products supplies extremely useful information to research, it also heightens concerns about privacy and data security.

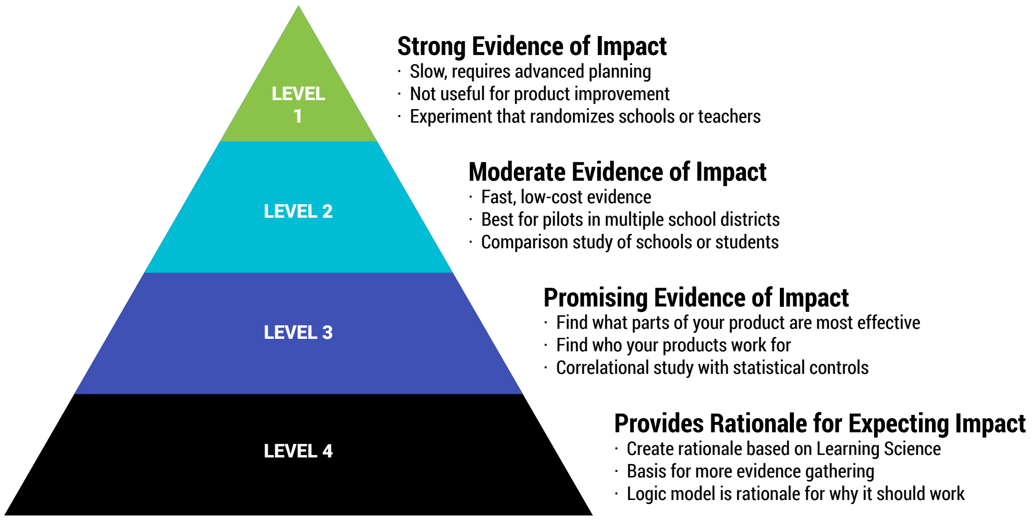
ESSA Evidence Standards

Recent federal legislation has put edtech research in an important new policy context. This is a major change that has occurred since the prior edition of these Guidelines, and its change in research policy provides a framework for many of the guidelines. Details of these standards are provided in Guidelines 6 through 9. Here we provide an overview of this important policy innovation.

The 2002 NCLB law brought "scientifically based research" into the middle of education policy. Following that, the Education Sciences Reform Act brought about the Institute of Education Sciences (IES), and following that, the establishment of the What Works Clearinghouse (WWC). WWC set strict standards for what were acceptable research designs to show that a program caused an impact. However, it wasn't until NCLB's successor, Every Student Succeeds Act (ESSA), was finally passed in late 2015 that a more encompassing, and usable, definition of evidence became available.

It is important to note that ESSA more clearly ties categories of federal funding to evidence. For example, states are required to set aside at least 7% of their Title I, Part A funds for school improvement programs that include "evidence-based" interventions meeting the highest three levels of evidence defined by ESSA. ESSA also adds "evidence-based" requirements for using Title II funds for class-size reduction and personalized professional development.

ESSA identifies four levels of evidence from the basic level (level 4) to what it considers strong (level 1) evidence, where each level allows for higher levels of funding from grants and other U.S. Department of Education (ED) programs. These levels are detailed in later Guidelines. Here we represent them conceptually as a triangle where the lowest level forms the base for any research that follows, and the apex, while considered in ESSA to be the strongest evidence, occupies a relatively small space, indicating its relative infrequent use in studies of edtech products. We expand on this framework in specific guidelines addressing the specific levels.



Strong Evidence of Impact
- Slow, requires advanced planning
- Not useful for product improvement
- Experiment that randomizes schools or teachers

Moderate Evidence of Impact
- Fast, low-cost evidence
- Best for pilots in multiple school districts
- Comparison study of schools or students

Promising Evidence of Impact
- Find what parts of your product are most effective
- Find who your products work for
- Correlational study with statistical controls

Provides Rationale for Expecting Impact
- Create rationale based on Learning Science
- Basis for more evidence gathering
- Logic model is rationale for why it should work

ESSA identifies four levels of evidence from the basic level (level 4) to strong (level 1) evidence, where each level allows for higher levels of funding from grants and other U.S. Department of Education programs.

ESSA presents the levels as a developmental sequence, where the early stages are prerequisites to later stages. The base (level 4) requires a rationale and a plan or commitment to a program of research. From that research, a provider may obtain evidence that a product is promising (level 3), which in turn warrants putting it to a more rigorous test in a study to obtain level 2 moderate evidence. Positive evidence from a moderately rigorous test justifies (but does not require) putting the product into a level 1 experimental test.

While ESSA is groundbreaking in its treatment of evidence, it is also early in its application, and there are still interpretations being worked out. Since legislative language is subject to interpretation, regulation, and even legal action, we will not pretend that the reporting provided in these Guidelines is definitive. It is a working hypothesis based on experience in the Investing in Innovation (i3) program, close review of the ESSA legislation and guidance, and conversations with other researchers. The law itself is somewhat cryptic, and definitions provided in the regulations are not entirely consistent with the guidance provided by ED. The authors of these Guidelines take our own path, adopting interpretations which we feel will best serve the developers and customers of edtech products and take full advantage of the innovative strengths of the new law.

Scope and Limitations

The Guidelines focus on the U.S. K-12 schools market where student and school outcome data are collected systematically, and where the standards defined in ESSA apply. Generally, the research principles behind the Guidelines may apply to other domains, including higher education, informal learning outside of school, products sold directly to parents or pre-K centers, and edtech products sold outside of the U.S. However, the examples and discussion in these Guidelines retain a U.S. K-12 focus.

The Guidelines also focus on a genre of research: evaluating the impact—that is, the effectiveness or efficacy (see Glossary)—of an edtech product on educational outcomes. Product impact research aims at comparing what happens with a new technology-based intervention to what would have happened if the intervention had not been introduced. Basing this work on the ESSA evidence standards, however, builds in a broader approach that includes setting goals, understanding how the product works, and how to improve its effectiveness.

The field of research known as "**learning science**" is important to ESSA's base level 4 and is very much tied to product design. But beyond pointing to its role in providing the rationale for why a product should have an impact, these Guidelines do not provide suggestions for conducting learning science, beyond the discussion in Guideline 1. So, our focus on evaluation of impact is not meant to diminish the value of other research genres to education providers and decision makers, including descriptive or laboratory studies (see Glossary). Other research purposes, methods, and designs are indeed important to guide product design, selection, and implementation. For example, cost-effectiveness research can usefully complement measures of effectiveness in guiding procurement decisions. Also, the provider's internal research has an important role in continuous improvement, development, testing, and refinement by asking what it was about the product that made it work, under what conditions it worked, and with whom it worked. Embracing the developmental progression in levels of evidence defined in ESSA and addressing these formative questions can help to fill out the logic model (see Glossary) and rationale for beginning impact evaluations of the product.

The Guidelines will not attempt to dictate standards for evaluation methodology. There are quite a few resources on evaluation methodology, of which we find the book by Will Shadish and collaborators to be the most useful. Here we attempt to flesh out how research can be relevant to educators, attuned to the current policy environment, and conducted using accepted methods. We also hope the document will help inform all stakeholders about the research challenges unique both to studies of technology and to provider-commissioned research in general. For specific guidance on designing, conducting, and reporting research, readers are encouraged to consult evaluators with experience in school settings.

The following sections elaborate on 16 Guidelines through discussion of background, rationale, and examples. They are organized into clusters representing the stages of Getting Started, Designing, Implementing, and Reporting.

THE GUIDELINES

I.

Getting Started

This first set of Guidelines starts with some very basic considerations and reminders of what to think and understand about the product and customers before engaging in research on the impact of a product. This includes getting started on the first stage in generating evidence of product impact, specifying how the product is supposed to have an impact, and determining what prior research to use in order to build the rationale for an expected impact.

1 Develop and document a model for how the product works

The base-level evidence considered by ESSA is a rationale for the product. Documenting this may involve a white paper and review of literature. The rationale is made explicit through a logic model or theory of action, which specifies the key components of the product, how it is implemented, the intermediate effects on teachers (if used in a classroom), and ultimately the impact on students (or other goal of the product). At each step, the model may specify how success can be measured. The hypothesized causal connections should be supported by prior research on similar products or learning processes.

Our first Guideline addresses the ESSA evidence level that sets the rationale for the product in terms of the learning science that it is based on. This is an essential first step in any program of research and represented earlier as the base of the evidence triangle. The base level of evidence (level 4) calls for a flow-chart diagram called a logic model. This is usually a diagram specifying the elements from left to right. In the regulations that ED put out for interpreting ESSA (found here), they offer the following definition:

"Logic model (also referred to as a theory of action) means a framework that identifies key project components of the proposed project (i.e., the active 'ingredients' that are hypothesized to be critical to achieving the relevant outcomes) and describes the theoretical and operational relationships among the key project components and relevant outcomes."

This model and any accompanying prior research conducted by others is what ESSA means by "demonstrates a rationale based on high quality research" ("high quality" is not explicitly defined). ESSA adds that if this level of evidence is to be used by an agency to acquire a product, it must also include "ongoing efforts to examine the effects" of the product. This level of evidence will justify, for example, a district using a state grant (or funding from a foundation that uses ESSA evidence criteria) to pilot a new product and to supply the provider with initial evidence of promise. It is interesting that this level requires ongoing research, suggesting it isn't enough to have a white paper on the topic. There is an expectation to at least conduct formative research (see Glossary). This level is adequate for local decisions and may be all that is needed for an entry-level grant.

The logic model-based rationale is an essential step toward gathering evidence and should continue to be improved as the product and its implementation are refined. Research design will continue to use it—a researcher will base the design and analysis of a randomized experiment (see Glossary) on a logic model and will often include the figure in the final research report.

The model for how the project works should be accompanied by the product's implementation plan, which can be much more detailed. If the product is not implemented properly, it is not likely to show an impact. The more explicit this implementation plan, the more likely the research will be able to show whether the implementation met the expectations in the model. The model should document, for both customers and researchers, these elements, as appropriate:

- the professional development required;
- the amount and type of technology infrastructure needed;
- the level and type of support the technology will require;
- the amount of time that should be devoted to each element of the intervention;
- a distinction between elements that are critical and elements that can be considered non-essential; and
- the appropriate curricular and instructional strategy within which the product or service was designed to work.

If multiple models are suitable or some elements are more important than others, these variances should be documented.

2 Know your customer—this is who provides the research sites

Research on edtech product impact takes place in schools and generally involves negotiating with school district decision makers for use of their data, access to teachers, and the like. This can make the process of finding a research site similar to the sales process. School districts (or charters or their management organizations) are generally the owners of the student and teacher records so are the most likely target of a search for research sites.

Depending on the kind of research, what you'll be looking for in a school district, charter management organization, or private school, as well as the process of negotiating with these entities, can be very different. Most often, it is a school district that has control over access to student data, so for convenience we will refer to this variety of agencies as a "district".

In many cases, there will be an opportunity to take advantage of a district's pilot of the product where the district has already purchased it for some, but not all, schools or teachers. Finding the right district is an art as well as a science. Schools may want to initiate a study to determine how well the product is working for them. Often, it is up to the provider or the researcher to identify a district to approach.

In the case of research that involves randomizing schools or classes to getting access to the product (or not), and potentially reaching level 1 of ESSA, it is necessary to find a district that has not yet implemented the program, or is using it in a small portion of classes. In this case, look for a district that would like to use the program and is probably somewhat dissatisfied with their current solution. In other words, an ideal district is also a prime target for a sale, which often introduces an awkward position of giving away the product for free to a school that could have been a customer, in exchange for the agreement to allow research to be conducted.

Finding the right district involves several other factors that are addressed in these Guidelines including how large a sample is needed, what kind of outcome measures will be available as part of the administrative data, and what solutions are currently in place (if any) to solve the problem that the product addresses.

Comparison Groups

A basic characteristic of impact research that falls into ESSA's level 1 or level 2 is that it compares a group using the new program to a group that is not. In general, we are measuring the difference between the existing baseline way of doing things and the performance of the new product intended to replace or supplement the existing approach. Thus, most impact research involves comparing a group of schools, teachers, or students using the new product to a group doing what they were going to do anyway, called “business as usual” (BAU). In education, there is no pure “double blind” experiment similar to a medical trial where the “treatment group” is given a medication and the “control group” is given a placebo, or sugar pill, with the subjects ignorant as to whether they are taking a real medication. In schools, a new math program is typically compared to the math program already in place. But here is the problem— while a placebo has zero effectiveness, the BAU has some, generally unknown, level of effectiveness. The same program may have a high impact when tested in a district with a weak BAU, and a negative impact (worse than the control group) when tested against a very strong BAU. If the study is conducted as a pilot of a new program in and for a specific district, this is not a problem since the BAU is exactly the program that the administrators intend the new program being piloted to replace. The administrator wants to know, does the new program do better than what’s currently in place in my district.

Where the point of the research is to generate evidence for use beyond the local administrator, a researcher or provider may prefer to recruit a district or districts where administrators are unhappy with their current BAU. While this may seem like gaming the system, it is very reasonable that schools which are very happy with the current solution are unlikely be good prospects for the product. For those looking to improve on what they are doing, their BAU group provides a valid comparison. It will also increase cooperation since the educators have a stake in whether it works for them. This is why it is both valid and productive to test the product where it could prove useful. At the same time, the impact demonstrated against a specific BAU, cannot generalize to another district’s BAU. This is why conducting multiple studies in different districts and with different populations will give a clearer picture of the conditions under which the product is most effective.

A Problem with Just Reporting Overall Impact

Comparison studies generally have as their primary finding the overall impact across the whole sample (“average treatment effect”). In educational research, where a placebo designed to have zero impact cannot be implemented, the size of the impact is determined as much by the quality of the comparison group or BAU as it is by the effectiveness of the new program being tested. As we’ve pointed out, this is not a problem where a district is conducting its own study and only cares about whether the new program is better than their current BAU. But if the purpose of the study report is to show the product’s usefulness to potential new customers, then the overall or average effect size found in the district is not informative. What is useful is information about the conditions that led to differences in the effect. For example, was it more effective for high achieving students or for students below grade level? Did experienced teachers find it more useful than novice teachers? ESSA (or the WWC rules that it uses to define its levels 1 and 2) is not concerned with the conditions that result in more or less impact. The WWC’s singular focus on the overall impact can cause a problem for decision-makers in the schools since they want to know whether the program is likely to work in their district given their resources, population, and the problems they need to address. They need to know about the subgroups in the study that got more or less benefit. The WWC does not prevent researchers from analyzing and reporting on differential impact, but their reviews of research do not report those findings. The school decision-makers will have to consult the original research report to get those findings.

Funding for Research

Finding the funding for studies will very often involve the school systems who are already customers or are potential prospects. Where research is paid for through the company's marketing or product groups, the work will require identifying a school system interested and willing to help. Government and foundation resources are available to school systems and may require that about 20% of the funding pay for an evaluation. Ultimately, any sale is an opportunity for collecting systematic data; and by offering schools a discount in exchange for collecting data, there can still be funds available for data analysis and reporting. Building on pilots paid for by the schools, the cost of research can be reduced.

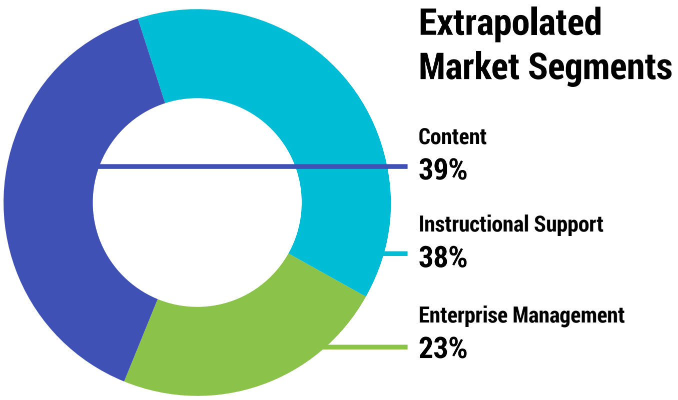
> By offering schools a discount in exchange for collecting data, there can still be funds available for data analysis and reporting.

3 Set a goal for your product that is measurable and realistic

In research on edtech, moving from the statement of the goals to finding a usable and technically acceptable outcome measure for the product raises several issues: (1) Many products are not intended to raise test scores, a common target of impact research; (2) It is important to use an outcome measure that matches the goals, but a test that is too closely aligned with the product goals may not be valid; (3) Providers should be realistic about the size of the difference their product will make.

Test Scores and Beyond

Technology provides a wide range of functions in schools. SIIA's 2014 Educational Technology Industry Market: PreK-12 report shows that 61% of the software industry revenue came from enterprise management and instructional support, such as student information systems, curriculum management systems, professional development programs, assessment applications, data warehousing systems, and information productivity applications. These are technologies that help make schools more efficient and productive. Instructional software used by students may still be the focus of accountability, but impact research can address the much broader scope of edtech.



Many options for outcome measures are becoming available including measures of social and emotional learning, school climate, teacher evaluations, teacher retention rates, classroom practice through student surveys, financial data, and quality of leadership. There has been a tendency to reduce the impact of any system to its effect on student test scores, the usual focus of accountability systems. A product's impact, however, can be measured using other outcomes for which the product is directly designed. For example, improvement in classroom discourse may not show up in the end of year test, but could affect student motivation and career choice that may pay dividends in the future. Impact research calls for measurements after a specified period of product use, so near-term indicators that predict future benefits must be used as the outcomes. Fortunately, many of these alternative measures are available in state or local administrative data, eliminating a need for expensive data collection as part of the research. Measures and their availability vary from one state and/or local agency to another but are sufficiently valid and reliable, and therefore suitable for use in research.

In building the logic model for the product, it is important to identify the kind of outcome that customers will expect after implementing the product. Identifying the specific measure may depend on where the research will be conducted, since different states and districts may collect different accountability measures. We should also note that most studies sponsored by ED focus on student outcomes and not, as the primary result of the study, other processes such as changes in classroom practices that may also be caused by the product's introduction. In a logic model, these often occupy an intermediate position and may be usefully measured as "mediators." The ESSA regulations talk about "relevant outcomes" that include outcomes other than those for students.

An Appropriate Balance between Sensitive and Meaningful

Using measures for which local educators are accountable will usually provide the most meaningful results for customers. In some cases, an outcome measure that is important for school accountability (e.g., state tests) may constitute too blunt an instrument to capture the full value of a certain product (e.g., one that is more narrowly targeted such that the state test includes content not covered by the intervention).

Another assessment may be better aligned to, and therefore better able to measure the impact of, a product with a narrower or more targeted goal—including achievement on a specific set of learning standards, student technology literacy, critical thinking, or student motivation.

Without going into a technical psychometric discussion, it is important to point out that considerable testing and statistical analysis go into the development of quality tests to ensure their validity and reliability. This can be a very expensive process. While developing a test specifically for a study may be necessary in some cases, doing so carries a risk of unreliable or invalid results because the test would not have gone through rounds of piloting and improvement. Even when designed well, such tests may do a good job of measuring the specific outcome but may not indicate whether the product or service is likely to make a difference on a more broadly-based measure (such as a high-stakes test). The provider should work with the researchers to identify outcome measures that will be meaningful metrics for the research audience, given the product or service being studied. The measure should not be over-aligned to the product or be developed as part of the product itself. In fact, reviewers will often reject studies where the outcome measure was developed by the developer. To capture both specific skills and broader constructs, more than one measure may be needed.

One important caution involves instances where the comparison group has no exposure to the concepts underlying the instructional product or service being studied. If the researcher administers a test to specifically assess those concepts, the results will have little meaning, and the size of the impact will be magnified by the lack of a "business as usual" to which the program can be compared. While such strategies can be used in laboratory research, reviewers of impact research, including the WWC, will reject a study based on a measure of content that only the treatment students had an opportunity to learn. Similarly, simple pretest-posttest studies without a comparison group are rejected as providing evidence of impact. They merely show that it is possible to teach the material in the test but give no sense of whether the product is any better than anything else at promoting that learning.

Important Impacts Can Often Appear to be Small Differences

The developer of a new product naturally hopes and expects that the innovation will make an enormous difference. Researchers, however, often find seemingly small differences. Research on impact always involves a comparison, e.g., users to non-users, or school outcomes in the years before and after an innovation is implemented. The difference being measured is between a new approach to achieving the goal and how the same goal was tackled in the past or in a comparison group. This is the basis for the objection to studies that do not provide a comparison (noted above).

The provider should work with the researchers to identify outcome measures that will be meaningful metrics for the research audience, given the product or service being studied. The measure should not be over-aligned to the product or be developed as part of the product itself.

The impact is often expressed in an "effect size" (see Glossary) where ESSA guidance has declared an effect size of 0.25 to be considered "substantively important." This is roughly equivalent to a 10-percentile difference. This may not sound huge, but in field research where new programs are being tested against business as usual, a difference of 0.1 is not uncommon and may make enough of a difference to warrant implementing a new product. Studies reporting effect sizes greater than 0.5 should be viewed with skepticism since the study may not have been conducted in a real field setting or may have made use of a test that was over-aligned with the product's goals.

We can derive two lessons from this situation. First, as we indicated earlier, the overall average impact, which is the only result considered in WWC reviews, may not be as important or useful for developers and for buyers as the different results found for different subgroups or conditions. Second, developers and buyers should not depend on a single study but should test the product under a variety of conditions in multiple districts. A single study, regardless of how high up in the ESSA hierarchy, cannot give a thumbs up or down on a product.

4 Consider how much support to offer during the product evaluation

Depending on the stage of the research program and product development, a provider may offer more or less implementation support to the program. In early stages, it is appropriate to assure that all users get whatever support is needed to be successful. However, when the goal is to generate evidence of impact, the level of support becomes relevant since support can translate into greater impact. Like support, time is also an important factor in getting an impact. Field studies in real schools take time for students to learn new skills or content, to change attitudes, or to see other effects such as acceptance to college. Likewise, for a teacher to improve classroom discourse or to figure out how to integrate technology, time and resources are needed.

Efficacy vs. Effectiveness

> In an efficacy study, the goal is often product improvement rather than, or in addition to, initial measurement of impact.… By contrast, an effectiveness study provides no more than the usual level of training and support that an ordinary customer would receive.

The term "efficacy research" is often used generically to refer to any research showing the impact a product is having or that helps improve the product's effectiveness. In more technical terminology, evaluation research is often divided into two general types. There is a distinction between *efficacy* studies that show how a product can work under ideal conditions and *effectiveness* studies that test it on a larger scale under regular field conditions.

In an efficacy study, the goal is often product improvement in addition to initial measurement of impact. In this case, the researcher (or provider), goes beyond the support and training ordinarily provided. This may include monitoring implementation and intervening to ensure delivery of the training, infrastructure, support, and anything else required to nurture the intervention and to ensure that it is implemented as recommended by the provider. Thus, efficacy studies show how a product or service works in the ideal case. An efficacy study is useful at the early stages of an intervention where the amount of support required is not certain, and it can be an excellent way to pilot research methods in preparation for a larger field study.

By contrast, an effectiveness study provides no more than the usual level of training and support that an ordinary customer would receive. In this kind of study, some, but likely not all, participants will implement the product as intended. While large impacts may be more difficult to obtain in an effectiveness study, the results may be more meaningful to schools, as they reflect ordinary conditions of implementation.

Intent to Treat Mode

Experimental researchers, especially those conducting randomized control trials (RCTs) designed to meet level 1 of ESSA evidence levels, can be quite strict in preventing the provider of the product being tested (the "treatment") from offering help that wouldn't normally be provided. So, in an RCT that is designated as an "intent to treat" (ITT) study (see Glossary), where a teacher, for example, doesn't attend the prescribed training, the results for their students are nevertheless included in calculating the treatment effect. The idea is that under normal conditions, some teachers will not make it to the training. The ITT study tries to measure the overall outcome, including the ordinary amount of non-compliance that schools might expect. Non-researchers, both the product's providers and their customers, may find this nonintuitive since they want to know how the program works when it is implemented as intended. That is, they assume that the research should be an efficacy study. This assumption may be reasonable since the specification of "normal" or "ordinary" may vary with resources available to the schools implementing the product.

Knowing that full-scale studies are usually conducted under the ITT assumption, it is important for the providers to be very clear about what the product implementation should include to be considered a fair instantiation of treatment. Some examples are: (1) How much ongoing coaching is considered support for the product? (2) How much does the success of the product rely on a majority of the teachers in the school working as a team, and how is that monitored and encouraged? (3) Are the district IT support staff expected to check in and make sure the network and devices are working as intended? All of these conditions for success can be identified in the logic model as necessary inputs along with the edtech software itself. In developing a product and preparing for research, providers need to explicitly state assumptions for how to achieve adequate implementation. The training, support, and monitoring needed for success has to be built into the normal delivery of the product.

Quality of Implementation

With edtech products, the usage data allows for precise measures of exposure and whether critical elements of the product were implemented. Providers often specify a specific amount of exposure or kind of usage is required to make a difference. And educators often want to know whether the program has an effect when implemented as intended. Researchers can readily use data generated by the product (usage metrics) to measure the kind and amount of implementation and identify users who used the program in the manner intended.

Since researchers generally track product implementation and statistical methods allow for adjustments for implementation differences, it is possible to estimate the impact on successful implementers. This can be especially useful in comparison studies designed to meet ESSA's level 2 moderate evidence. It is, however, very important that the criteria researchers use in setting a threshold be grounded in a model of how the program works. Guideline 1 explains that the first step in conducting impact research is to develop and document a theory of action. This will, for example, point to critical components that can be referred to in specifying what counts as adequate implementation. Without a clear rationale for the threshold set in advance, the researcher may appear to be "fishing" for the amount of usage that produces an effect.

Some researchers reject comparison studies where identification of the treatment group occurs after the product implementation has begun. This is based on the concern that the subset of users, especially users who comply with the suggested implementation, may differ in motivation or other personal characteristics from non-users or non-compliers. While after-the-fact identification of the treatment group can present certain problems, these guidelines take the position that identifying a treatment group on the basis of usage of the product or on the basis of a pre-specified level of usage is also a good design to meet ESSA's level 2. In this design, it is essential that the level of implementation is clearly described in the report and has a solid foundation in the product's theory of action or past studies of this or similar products.

Time Needed in Field Studies

Duration is an important consideration for any study design. Generally, the longer the product or service is used, the more likely its effects will be measurable. There are two areas of concern with duration: (1) reaching full implementation of the product or service and (2) providing sufficient exposure, for both teachers and students, to the product or service once it is fully implemented.

Some educational technologies are fully implemented very quickly because, for example, professional development requirements are minimal or a technology is readily put into use. In other cases, it is often not possible, even with extraordinary effort, to get an intervention up to speed in the first months of a study. Some are designed to be rolled out over time. For others, it is recognized that the professional development takes time for participants to absorb. In cases where implementation takes longer, having the study extend over two or three years is not unreasonable, and an interim report may focus more on implementation than on outcomes.

The second duration issue is the length of time before a product or service, once fully implemented, is expected to show an impact. When students move through the educational content slowly, a full year of implementation is likely needed. It is also possible to envision a study in two phases (e.g., two semesters or two school years), perhaps each with a pretest and posttest. In this way, interim results can be reported prior to the full report. Similarly, with studies using data from prior years, choosing sites that have been using the product or service for two years or more may provide stronger results where one year is not sufficient time for full implementation.

The Laboratory Mode

In contrast to effectiveness (with its ITT extreme) and many efficacy studies (which we consider field research), laboratory studies are tests of small interventions over short periods of time and studies of learning processes. Laboratory research may be conducted in a single classroom with students interacting with a research assistant or with a teacher trying out a unit and pre- and post-testing that content. Here, the content to be learned is quite specific and limited in scale, so the length of the study can be much shorter. A concern with a very short study is that the impact of the intervention may be greatly overestimated. In contrast to a field study, the extra support by the researcher to assure that the treatment is fully inserted into the classroom and the general excitement of trying something new may inflate the results. Also of concern with laboratory studies is the match between the treatment and outcome measure—an impact may not register if the test is much larger in scope than that addressed by the intervention—so often the outcome is much more closely aligned to what was taught than would be acceptable in field research.

This example illustrates how the research can affect the normal ecology of the school and shows that the research design should be carefully attuned to the ingredients that make for a productive product implementation.

As we turn now to potential issues in research designs, the provider of edtech products should consider how the ways the study is put together may inhibit important elements of the implementation assumed in the logic model. Here is one example to keep in mind as we address the issues of research design. Consider an edtech product that would normally be adopted by a district to supplement middle school math instruction and be supported by district staff, involving team meetings in each school to integrate it into the existing curriculum. If the research design calls for some teachers in the school to use the product and others to continue with "business as usual," the dynamic within the school may be affected. Looking for a meeting place away from the staff in the control group may prove a bit awkward and the principal may be hesitant to generate enthusiasm for the innovation. This example illustrates how the research can affect the normal ecology of the school and shows that the research design should be carefully attuned to the ingredients that make for a productive product implementation.

II.

Designing the Research

This second set of guidelines focuses on the types of research designs that make the difference between what does and does not count as evidence for ESSA. These Guidelines are not meant to be a training course for researchers, but they should introduce the costs—in time and resources—and better prepare you to engage researchers in studies that will meet the needs of your organization. We also hope that these guidelines will supply researchers and edtech providers with a common language so they can get beyond the technical jargon of the evaluation research world.

5 Decide who is being tested: students, teachers, schools, or a combination?

Identifying the study's *unit of implementation* and *unit of analysis* is fundamental to how products are tested and what conclusions to draw. Unit of implementation refers to the scale at which the program is adopted: classroom, school, district, etc. Unit of analysis is the level of granularity of outcomes: if researchers use individual student test results to assess program effect, school averages, or teacher performance metrics. Determining the unit of analysis—essential for deciding on the size of the study sample and design of a study—follows somewhat from the model of implementation. One way to think about it is to ask: "What is the smallest unit that can independently work with the product?" The considerations include how the product will be used, what analytic strategies are appropriate, and how large a study will be needed to see a significant effect.

Level of Implementation and Analysis

Here are some examples of units of implementation and analysis. If the product is a whole-school innovation—e.g. a blended learning system—then the unit of implementation is the school. When measuring impact on individual students, such as their test scores, the unit of analysis is the student. When looking at percentages of students that meet reading standards by school, the school is the unit of analysis. In either case, it is assumed that overall improvement of the school is the goal and students are considered to be *clustered* within a school, that is, some common factors or school characteristics influence the performance of all students in each school. In contrast, a product that personalizes instruction is delivered directly to students, who create individual user accounts. This may put both the unit of analysis and the unit of implementation at the student level.

For a home-based tutoring system sold on a consumer basis, the smallest a unit may be is an individual student. However, most edtech products are designed to be used in organizational settings, such as a teacher using the product personalized for all students in a classroom. Formative assessment systems may be designed for a whole school and may call for leadership training and schoolwide support for implementation. Course management systems may be implemented districtwide. In all cases, data may be collected on individual students or teachers (the unit of analysis), and data are analyzed using those individual data points, but the data points are clustered at the unit of implementation in the analysis.

The level of implementation will help determine the size and cost of a study. A non-technical way to think about it is that the number of units needed for an experiment is similar whether the unit is an individual student, a class, a grade level, a school, or a district. For example, if the technology is implemented across a whole school, as in a school-wide reform, we may need 40 schools (and many thousands of students) for the study to detect a difference in test scores of students in program schools compared to students in control schools. If the intervention is provided to each student independently, as in an instructional software accessed via a login with students assigned at random to use the product or not, we may need fewer than 100 students. Now 100 students is a larger number than 40 schools, but the contrast to be considered is the 100 students compared to the thousands of students needed in the school-level study. The larger units will naturally involve far more individuals—and cost—than the smaller units.

Determining the Size of an Experiment

> The larger the expected effect, the smaller the experiment (in number of units) needed to detect the difference.

These guidelines will not attempt to provide ballpark numbers of the correct size of an experiment. There is an important and complex process by which a researcher makes this estimate. Many factors go into determining the number of units needed in an experiment. The researcher may ask the provider how big an effect they think the product will have (and may study the research literature to find the size of impact of similar products). The larger the expected effect, the smaller the experiment (in number of units) needed to detect the difference. Knowing how the student, school, or other cluster did before the study on the measure that is used as the outcome is essential in any impact studies. The pretest greatly improves the precision of the estimates of impact from the study and reduces the number of units needed.

There are other technical factors such as the *intraclass correlation coefficient* or ICC, the explanation of which is beyond the scope of these Guidelines. Suffice to say, ICC addresses the effect of clustering on the precision of the impact calculation. Where the ICC is calculated incorrectly (or ignored), the experiment may incorrectly claim statistical significance (see Glossary).

In some cases, the implementation model can adapt to keep study costs down. For example, although a formative assessment system would ideally be implemented school-wide, for an effectiveness study, it may be more efficient to implement the system at some grades and not at others. This increases the number of implementation units (grades instead of whole schools) given the same number of schools. Similarly, for the purposes of an effectiveness comparison, a technology-enhanced curriculum that would normally be used by a teacher in all sections of a course might instead be implemented in half of the teacher's class periods, while the other classes might continue working with the existing program. Instead of working with six algebra teachers, each with 100 students, we have 24 class period units, each with 25 students. This change in design can generate a substantial increase in power to detect differences between the students with the new algebra product vs. the business as usual textbook.

There are two related arguments against dividing up the normal unit of implementation. First, there is a danger of what researchers call “contamination.” For example, when a teacher splits up class periods, it is likely that at least some of the techniques the teacher learned in the context of the treatment program will carry over into the comparison classrooms. The consequence of this contamination is that the comparison group students get some of the advantage of the product, thus reducing the apparent effectiveness of the product or service under study.

Second, dividing up the normal unit of implementation could disrupt the normal collaboration or support systems in the school, and the product or service might not perform as well as it would otherwise. Similarly, if only a few teachers in each district use the intervention, as part of a larger multi-district experiment, the administrative/technical support, training, or leadership needed for implementation may be insufficient. Therefore, when implementation depends on resources and leadership at the school or district level, or when collaboration among a group of teachers facilitates the integration of a product or service, it may be counterproductive to design an experiment in which teachers use the technology in isolation.

Looking at Different Factors like Subgroup Characteristics

If the goal is to identify the conditions under which the product has the strongest effect, it is often necessary to look at whether there is a differential effect for different subgroups of students or teachers. The conventional approach to analysis and reporting tends to discourage looking at subgroup impacts because the researcher may appear to have been fishing for a positive result to report. For example, after running a study a researcher may want to explore the data by looking at 20 different subgroups. This can be useful to get a picture of whether there are some groups with a much larger impact than others, but given that with 20 comparisons, one is likely to appear significant just by chance. So, there’s a general prohibition against considering that result as something confirmed by the study.

As the field moves toward a greater concern with subgroup effects, there are two recommendations that are being proposed for getting around the perception of fishing for results. The first is for the research community to agree on a set of a dozen or so factors (subgroups, conditions, etc.) that will be analyzed in all studies so that the accusation of fishing can’t be raised and so that studies can be compared with each other to find recurring patterns. The second is for the larger community of providers, buyers, and researchers to move away from the idea that a single study is enough to prove a product effective or not. Instead, the routine expectation should be that studies of each product are conducted under a variety of conditions, with analytical technique called meta-analysis used to examine impacts for different subgroups across studies.

6 Consider the four levels of evidence defined in ESSA

ESSA provides relatively clear definitions of the four levels of evidence of product impact. We say "relatively" as compared to prior legislation such as NCLB, and do not mean to imply the levels are without ambiguity and debate as to their interpretation. These Guidelines interpret the four levels of evidence as a developmental progression of increasing sophistication and rigor in showing that the product does (or is likely to) have an impact on the desired outcomes. In the ESSA legislation, this sequence is tied to eligibility for increasingly more generous funding.

The levels of evidence and levels of funding were "piloted" in the work within ED, particularly in the rules for awarding i3 grants and adopted under the ESSA-aligned Education Innovation and Research (EIR) program. Each grant included the requirement for an evaluation aimed at a level higher than the level of evidence already established for the program. Programs with stronger evidence were eligible for larger grant amounts that would fund larger scale and more rigorous studies. As mentioned in the introduction, this hierarchy was built into the evidence standards that became law with the signing of ESSA in December 2015.

As previously mentioned, it is important to keep in mind that ESSA influences how some—but not all—federal funds administered by ED that flow down to the states are awarded. State and local agencies are encouraged to use the same definitions. Interactions with local decision makers may relate to factors other than the strength of the evidence, but where evidence is asked for, the levels defined in ESSA provide a user-friendly hierarchy. This hierarchy ranges from the rationale based on learning science—which does not test whether the product is effective but gives a rationale for why it is worth trying—to what is sometimes called "proven" effectiveness based on multiple high-quality experiments.

The ESSA rules for levels of evidence are fundamentally aimed at the design of the study without regard for whether it works as well for all subpopulations or under different conditions. The research design requirements ensure that the study shows that the product caused the outcome, as opposed to the outcome being caused by another factor that happens to be correlated with both the product and the outcome. For example, better trained teachers, higher performing students, or more savvy principals may be more likely to adopt or be given a tech product, thereby confounding the effects of the product with those of the characteristics of the users. The goal of the research methods specified in ESSA is to remove the influence of competing causal factors (called confounding factors – see Glossary) so that the intervention under study can be credited with any observed changes.

ESSA defines four conditions for the application of "evidence-based" to an "activity, strategy, or intervention." There are additional criteria, e.g., concerning the size of the sample needed, but we will focus on required design characteristics.

Strong Evidence

As we describe in Guideline 7, this level calls for at least one experimental study that uses randomization where participants (e.g., schools or teachers) are assigned at random to receiving the program or being in the control group. Random assignment allows us to expect (although not definitively) the two groups to be equivalent, and for that reason, the WWC accepts the design "without reservations" to demonstrate that in the experiment, the program caused an impact.

Moderate Evidence

Level 2 calls for least one "quasi-experimental study" (see Glossary). A quasi-experiment is a comparison study where a group that has used the program is compared to another group that has not used the program, and where both groups have very similar characteristics, (the program is not assigned to the groups at random). We address this in Guideline 8.

Promising Evidence

Level 3 calls for a "correlational study with controls for selection bias" (see Glossary). This is new and is not acceptable by WWC standards. However, it gives legitimacy to research designs that not only provide initial evidence that the product may be effective, but also provide developers with rich information about how the product works. "Promising" is the key. It means that while the study provides weak evidence of impact, it nevertheless indicates that the product is worth exploring further. We address this in Guideline 9.

Base Level

Level 4 was addressed in Guideline 1. It does not involve a study of the program in question beyond the planning stage, but provides justification for trying it out and evaluating it for evidence of impact. We call this the "base level" because it gives a provider a place to start. Level 4 calls for providing a rationale based on learning science that shows that strategies similar to the program being considered are likely to improve relevant outcomes. At this level, there is a research requirement, which is to cite "ongoing efforts" to study the program. This allows a school system to participate in an ongoing program of research prior to the availability of correlational findings.

7 Use random assignment if you have control over who gets the program now and who gets it later

Studies that randomly assign units (e.g., teachers, schools) to using the product or business as usual are often called randomized control trials (RCTs) or randomized experiments. Randomization assures that the participant's preferences do not play a role in whether they are assigned to use the product versus assigned to follow business as usual. An RCT requires more preplanning than a comparison study in which users are not randomly selected, but once set up is generally simpler to design and easier for the researcher to analyze.

ESSA's Level 1

The apex of the ESSA evidence standards—level 1: strong evidence—involves random assignment of units (e.g., teachers, schools) to using the product or not. There is a mystique about the randomized control trial (RCT), such that it is believed to be very expensive and take many years, yet is the unassailable "gold standard" for impact research. None of these things are necessarily true, although ED has been known to spend millions of dollars on RCTs and years reviewing them before finally releasing a report. This is not because RCTs are inherently expensive, but because large-scale government contracts are often attempting to answer a lot of questions in a lot of contexts in one study, and therefore can get caught up in bureaucratic complexity that may not be characteristic of all funding sources. For example, we have seen foundations funding "low-cost RCTs," using school district administrative data and opportunistic pilots of products already slated for adoption by the district. ED's Institute of Education Sciences (IES) is also moving in this direction with research programs emphasizing the opportunistic approach. It is useful to note that an RCT is generally simpler to design, requires less information, and is easier for the researcher to analyze, all of which should ultimately help to lower the cost.

Random assignment of units to conditions has an enormous advantage in eliminating sources of selection bias (see Glossary). It eliminates most other plausible explanations by providing teachers or schools with the product based entirely on the result of a random coin toss (or some equivalent), rather than through personal choice or confounding factors. Interest and morale play no role in the assignment of the schools to use the new technology. The researcher's estimate of the technology's impact is not biased by these or any other characteristics, although the two groups may still be uneven on some factors just by chance. The units (teachers, schools, etc.) do not choose whether to join the program. So, all the characteristics, whether the researchers know about them or not, that otherwise would have led some participants to adopt the program and others not, do not influence the results. This is why RCTs are referred to as the "gold standard"; there's no better way of avoiding the bias resulting from the participants being able to choose whether to use the product. However, there are still things that can go wrong. While challenges can occur with any research design, there are some issues unique to RCTs.

Bias within an RCT

Although highly valued by researchers for eliminating important sources of bias, RCTs are not perfect.

Prior to randomization, voluntary participation in the study causes the sample to have its own kind of selection bias. Participants, i.e., units of randomization (whether schools, teachers, etc.) are likely to be less risk averse, interested in the product, and generally more enthusiastic than the rest of the population. If a school district adopts the product for everybody after getting good results in the teacher-randomized RCT, it may not work as well among the less motivated teacher population as it did among those who volunteered to participate.

Other problems that can occur when some teachers don't get the product include teachers becoming demoralized (although we seldom see that among RCT volunteers), comparison group teachers competing with the treatment group, or comparison teachers adopting a new product. These possibilities are seldom measured or reported.

Some Challenges in Conducting RCTs

While not necessarily more expensive or time consuming than other kinds of studies, RCTs may present unique challenges.

Need for planning ahead. A major impediment to random assignment in schools is that it requires concurrent planning of the evaluation and the rollout of the intervention. Very often, administrators at the research site have already promised the intervention to some schools before researchers can influence the method of assignment. The need for advanced planning could be a major reason RCTs don't happen.

Deprivation of the control group. Some argue that group assignment deprives some (equally needy) students of the new product or service. There are a variety of approaches to address this complaint. Using a lottery to assign participants may be the fairest method in cases where there is limited availability of the intervention or where it is being rolled out in phases. Some methods ensure equality by providing the control group with the intervention either after the study is completed or in a later stage of the study. One can also argue that, because it is unknown whether the impact will be positive, the control group is not necessarily being deprived of a benefit.

Need to start from scratch. A difficulty with RCTs is that in most cases, none of the participants are allowed to have used the product before. Random assignment is among participants who have not already chosen to use the product. While other research designs can use retrospective data from a product implementation and capture several years of use, the RCT participants must be using the product (or implementing it in the particular manner being tested) for the first time. A reason to keep an RCT running for a second year is to give participants more time to learn to be successful with the product.

Attrition. RCTs are sensitive to losing participants, especially if more are lost from one group than from the other, causing the two groups to no longer be equivalent. This is a problem with multi-year trials, particularly with principals and teachers changing jobs or going on leave, and with students moving to different districts. If differential attrition (more lost from one condition or the other) occurs because, for example, the product was unpopular, or because control teachers were dissatisfied, then the RCT is biased. Reviewers may downgrade an RCT to level 2 or level 3 evidence.

Calling the Shot

Researchers must be able to specify where the impact is expected to be found and why. This must be done before the outcome data are analyzed and should already be indicated in the product's logic model.

A final point that follows the important exploratory research activities and applies across evidence levels above the base level 4: researchers must be able to specify where the impact is expected to be found and why. This must be done before the outcome data are analyzed and should already be indicated in the product's logic model (Guideline 1). Ideally, when planning a study, researchers identify one or two outcomes that they believe are most likely to be impacted by (or that will moderate the impact of) the intervention. Researchers then commit to limiting their firm conclusions to those outcomes. As the number of observed outcomes increases, the chances increase of finding, just by chance, at least one statistically significant conclusion about the impact of the product or service; that is, to mistakenly attribute a chance difference to the effect of the intervention. For example, a test may report results in terms of five separate measures or subscales. If all subscales were treated as equally important, the likelihood that one would appear to measure an impact just by chance would be greater than if the most relevant subscale were identified initially as having more importance. Declaring a limited number of primary outcomes at the start of the study allows greater confidence that researchers are not mistaking chance effects for true effects.

8 Use comparison group studies to show evidence of impact

Studies using comparison group designs, often called quasi-experiments, can show evidence of impact by demonstrating the difference between a group that used the product and one that did not, i.e., engaged in business as usual. (RCTs that qualify for level 1 are also comparison group designs but use randomization rather than matching to identify the comparison.) This is ESSA level 2: moderate evidence.

ESSA's Level 2

The law considers these kinds of designs to provide moderate evidence of impact. The law itself declares that a product is evidence-based if it is:

> "…an activity, strategy, or intervention that demonstrates a statistically significant effect on improving student outcomes or other relevant outcomes based on moderate evidence from at least 1 well-designed and well-implemented quasi-experimental study."

The regulatory guidance goes into more detail specifying that moderate evidence of effectiveness requires a quasi-experiment as defined by ED's WWC, which results in meeting their standards "with reservations." The regulations also specify that a study cannot be considered evidence for product impact unless the study was conducted with a population similar to that of the district planning to acquire the product.

The WWC—established 15 years ago—has very detailed rules for approval of studies, and the regulatory guidance adopts them in their definition of the top two levels of ESSA's evidence-based standards. The reason the WWC has reservations about comparison studies is basically the problem of selection bias. If the students, teachers, schools, districts, or states self-selected the product, we can never know if that choice was driven by a characteristic that (a) we have not measured, and (b) results in a favorable outcome. If our comparison group doesn't have that characteristic that the program group has, and for that reason, doesn't have a favorable outcome, we are mistaking the product rather than the unmeasured characteristic as the cause of the difference in performance.

A common complaint about poorly designed research is that it inadvertently stacks the deck in favor of the intervention. For example, if schools or teachers with the most interest are chosen to pilot the product or service, the results could be biased. Their interest and enthusiasm as volunteers or early implementers may accompany other strengths that could reasonably provide an alternative explanation for differing results.

In another example, when schools or students who most need the intervention receive it, a comparison to other schools within their district is difficult. Even comparing the intervention schools with similar schools in other districts introduces a potential bias, in that the many possible discrepancies between the districts could account for any differences in performance found between their schools.

Nevertheless, a quasi-experiment can be quite useful. Generally, researchers attempt to identify productive variables—those likely to be related to an individual's propensity to select into the product user group. Many statistical techniques have been developed to match user units with non-user units. None are perfect, but it is reasonable to say that they produce moderate evidence.

Variations of the Quasi-Experimental (QE) Design

The basic idea is that the comparison group closely represents what the group using the product would have achieved without the new product. So, it is necessary that the two groups are as similar as possible before the product is introduced, that is, at baseline. Most critically, the groups have to be equivalent on the measure that will be used as the outcome, referred to as the "pretest." ESSA (and WWC) have set a standard for similarity: for any characteristic, especially the pretest, the two groups can't be more than a ¼ of a standard deviation (see Glossary) apart. In less technical language, that would be "reasonably similar."

It is particularly important to identify the user and comparison groups prior to examining the data on outcomes. Some reviewers will reject a study where groups are identified retroactively, that is, after the outcomes are known or knowable. A researcher may be tempted to use the outcome information to stack the deck in selecting groups.

In Guideline 5, we talked about the unit of product implementation (individual students, teachers, grade level teams, whole schools, or whole districts). Since these are the units to be compared, matching characteristics at that level will be particularly important. In a school-level product implementation, select other schools with similar characteristics within the same district or state. Characteristics of the school may include aggregate information about the students or teachers, such as the percent of the students in the federal Free and Reduced-Price Lunch (FRPL) program, or the average ratings of teacher effectiveness. In a teacher-level implementation, teacher characteristics—such as the average achievement of their students in the prior year—could be a critical matching variable. Surveys of teachers throughout the whole district (not just those who adopted the product) may be an important supplement to the district's administrative data in identifying a well-matched comparison group.

Generally, the comparison group is using the business-as-usual product that is already in place. Thus, the size of the impact of the intervention being tested is always relative to the effectiveness of the other, often pre-existing, program. The question isn't simply how much growth occurred in the user group; instead, we ask how much more growth occurred in the user group compared to what happened in the other group. Since we are generally comparing the product to what would otherwise be going on (the "business as usual" condition), it certainly helps to know what the condition consists of. In some studies, a competing product is already in place, so the study involves a head-to-head comparison. Sometimes, the product is up against cases where multiple products are in use. This may require surveys of the teachers, which will add to the cost of an evaluation.

QEs can be conducted when only aggregate (school average, frequency, or percentage) data (see Glossary) are available. There is an advantage, for example, in getting average data for a school in that it avoids personally identifiable information, or PII (see Glossary). We discuss data privacy later in Guideline 10. Caution is needed when interpreting the size of the effect, because it will apply only to the effect on the whole school (such as increase in percent proficient), as opposed to an impact on individual students.

Consecutive cohorts. A comparison group can be based on the prior year's results. For example, last year's sixth grade students (using a product that was already in place) are compared to this year's sixth grade students (using the product being studied). Note that this is different from comparing the pretest and posttest scores for a single group of students, which measures growth of that cohort but doesn't compare their growth to students who didn't use the program. Here, we are comparing a group with the program to one without, and for each cohort, pretest and posttest scores are needed. This is called a consecutive cohorts design and can be useful if other explanations for changes, such as environmental or social conditions at the research site, can be eliminated. Because it does not control for other differences that may occur from year to year, this design is not accepted by the WWC, so reviewers may not view this as meeting ESSA level 2.

Here is an example of the considerations in using this approach. The principal of a school may observe a good result—for example, the school's percentage of proficient students in a particular grade increased over the prior year—after having implemented an educational technology. The researcher would ask whether other things also happened that could explain the improvement. Some questions may include: (1) Were there other changes to the curriculum or instruction? (2) Was the test rescaled? (3) Did new teachers join the school? (4) Did the boundaries of the school neighborhood change? These and many other questions are quite reasonable and, as a practical matter for a local decision, the principal may know that the answer to all is "no." From the principal's point of view, it is not unreasonable to conclude that the technology is a likely explanation for the improvement.

Virtual Control Group. An approach that does not meet the requirement of addressing selection bias is often referred to as a "virtual control group" (or VCG). This approach depends on having a large database of student test results, where demographic characteristics are known for each student and where the test is based on a growth scale. For all students who used an edtech product during the year (or semester), the researcher obtains their pretest and posttest scores, (i.e., growth scores) on the same scale as those compiled in the database. These students are then matched to students from the database, based on the demographics known about both groups of students. Essentially, this design compares the outcomes of the user group to the normal average growth of other students with the same demographic characteristics. While this method does match on baseline pretest, it does not make use of information about what the comparison students are doing instead of using the product. And often it does not include information on the specific characteristics of the schools or teachers that would account for the students being in the user group, that is, controlling for selection bias.

9 Use correlational designs to find a program's promise and how usage relates to outcomes of interest

Studies using correlational designs can show that the product has promise but cannot really show evidence of impact. However, promise is enough to warrant small investments and can be the rationale for conducting a more rigorous study, such as a pilot in a customer district. This is ESSA level 3: evidence of promise.

ESSA's Level 3

The law describes an evidence-based product at this level as:

> "…an activity, strategy, or intervention that demonstrates a statistically significant effect on improving student outcomes or other relevant outcomes based on promising evidence from at least 1 well designed and well-implemented correlational study with statistical controls for selection bias."

This raises many questions. First, what is a "correlational study"? The official regulations (excerpted) for implementing ESSA help somewhat.

> ***"Evidence of promise*** means there is empirical evidence to support the theoretical linkage(s) between at least one critical component and at least one relevant outcome presented in the logic model for the proposed process, product, strategy, or practice. Specifically, evidence of promise [is based on] at least one… correlational study with statistical controls for selection bias [which] found a statistically significant or substantively important (defined as a difference of 0.25 standard deviations or larger) favorable [linkage(s)]."

This definition says the correlation is between a critical component of the logic model (see Guideline 1) and an outcome. This can't be any aspect of the product except something that is thought to be central to how the product works. In edtech products, this is usually an aspect of how the product is used, about which data are collected. For example, one might theorize that reading stories presented in a product leads to improved reading scores. A study linking student usage (e.g. the number of stories read) to reading test scores—and finding a positive correlation between the two—can be a demonstration of promise.

Ways to Use Correlational Studies

The statistical techniques used in correlational analysis can be more involved than just seeing if there's an association between two variables. Often, regression methods (see Glossary) that control for the influence of multiple variables are used and make the approach more useful.

- A correlational study can be done with data just from the students (or other units) that are using the product. The question answered by the study is whether the outcome of interest is correlated by an important aspect of usage. Since there is no need for a comparison group, conducting the study is much simpler. Of course, the study must control for selection bias to be acceptable under the ESSA definition.
- Looking at correlations (with appropriate statistical controls) within a comparison study can show linkages and corroborate and help to provide an explanation for the impact (as predicted by the logic model).
- It can be useful to check for correlations between implementation and outcomes, for the purposes of product improvement and understanding best practices. In many studies, even with the concerted effort of staff support and training, the implementation is variable, and deviations from the ideal model are found. Looking at which elements of the product were related to impact can be useful.
- By looking at the most successful users instead of just the average case, it is possible to identify best practices or to identify support and training events that were related to strong outcomes. Such information may help to shape product development, implementation models, and implementation itself.

After-the-fact exploration of the results for these kinds of linkages can also help shape the next round of research, when more refined hypotheses can be tested.

A Broader Understanding of Correlation

Looking for the relationships between two variables is a basic part of any statistical analysis. In an experiment, the researcher may be looking for a relationship between the student using the program or not (in the "treatment group" or not) and the outcome. This opens the reasonable approach of accepting research as valid at level 3 that failed on a technicality at level 1 or 2. For example, if a researcher conducts a level 2 QE but doesn't find a significant difference, and then examines only teachers who implemented the product as intended and compares them to matching teachers and finds a significant difference, this may be counted as showing promise, level 3, even though it did not succeed at level 2. Level 3, thus, becomes a broad category, where studies rejected by WWC—or at the equivalent ESSA levels—can find value and justify modest funding for a school district or investment in a product.

III.

Implementing the Design

The third set of guidelines focuses on actual implementation of the research design. The researcher may have devised an excellent design with clear research questions, sensitive measures, and clearly-differentiated comparisons, but now he or she must make it happen. The researcher may say that the study is designed to meet WWC standards without reservation, but, for example, if too many participants drop out of the study, the research will not be accepted by WWC. In this section, we address a set of guidelines about research "operations."

10 Use caution in handling confidential information especially personally identifiable information (PII)

The issues around getting data for research have been made more sensitive by the controversy and public concern about online privacy and the possibility of providers using student profile information for commercial purposes. However, cloud-based usage data can be considered essential in research on edtech products.

Access to PII for Research

Note that aggregate school data (e.g., averages for schools broken out by grade levels and demographic characteristics) can be obtained from state websites and other sources.

With cloud-based edtech, where information about students and other users is collected, researchers and providers must be particularly cautious in handling PII and other kinds of confidential information. School districts (or charter management organizations) control and are responsible for data on students and others. Researchers will need a data sharing agreement with the district before research using PII starts. SIIA's series of reports on Student Data Privacy & Security Laws provides a useful review of legislation including variations from state to state. It is important to note that exceptions are often made for using data in research.

Note that aggregate school data (e.g., averages for schools broken out by grade levels and demographic characteristics) can be obtained from state websites and other sources. Most, but not all, edtech providers can also give a research organization aggregate data through records of product usage, but this depends on the details in the provider's contract with districts.

The Family Educational Rights and Privacy Act (FERPA) sets the conditions under which schools can provide student record data to third parties, such as research organizations, without explicit parental consent. A requirement for parental permission for student data may be impractical and a cost impediment when large numbers of students and schools are involved. In short, FERPA allows a school district to provide data without explicit parental consent when the purpose of the research is to improve education and where the identity of the students remains confidential (is not released or exposed by the researcher). If the goal of the research is to simply generate marketing statements (which provides no educational value to the school system), it may fail this important provision.

Many states and school districts have their own procedures that go beyond the FERPA requirements. While the authority to release confidential student-level records resides in the school district, state databases are becoming an increasingly accessible source for very detailed school records that do not include personally identifiable information. [Note: These regulations are complicated, and neither this summary nor anything else in these guidelines should be viewed as legal guidance!]

The Institutional Review Board (IRB) Review Regarding Harm to Research "Subjects"

Research organizations, whether academic, nonprofit, or for-profit, must obtain approval of an IRB (see Glossary). The IRB reviews proposed research procedures and determines whether the activity constitutes legitimate research, whether there is a plan to obtain consent when required, and whether there is a risk of harm to participants. Most large organizations and universities have their own IRB formed in their own staff.

Large school districts with their own in-staff IRB may call for a review before research is conducted. Smaller research organizations often contract with companies that provide IRB services. Different IRBs vary in how strictly they interpret the basic requirements.

Any assessment of impact is considered research since the intention is to claim that the product will also work in other settings.

The first question an IRB will ask is whether the activity is "research," meaning that there is intent to generalize the findings. Purely formative studies or product testing aimed at improving the product is not considered research by this definition. Any assessment of impact is considered research since the intention is to claim that the product will also work in other settings. The next question is: Is it "human subjects research?" Human subjects research involves people that the researcher can identify. The IRB review can be simpler if only de-identified data are used.

IRBs are concerned about any risk to the participants, and if there is a risk, participants should knowingly consent, usually in writing. Consent is generally required if the research is going to do anything outside of the ordinary. For most of the impact research conducted in schools, the use of the product falls within the normal expected educational activities, and there is no appreciable risk to students or teachers. The assumption is that teachers and school administration are free to introduce new teaching tools and assessments including pilot tests of products. Consent is required when the research includes surveying teachers, interviewing students, and collecting video. This is especially the case where information is sought on sensitive topics like drug use or sexual activity. Parental consent, either active (an actual signature) or passive (responding to a notice sent home if there's an objection) is sometimes required depending on the activity. In many cases, the potential harm to the participant would come from inadvertent release of the data with PII.

An important principle in experimental research is that participation must be voluntary. An IRB will generally enforce the idea that it is unethical to allow a supervisor to compel participation of a teacher. Supervisors can perhaps require the use of mandated products and services, but not participation in an experiment. In a randomized experiment, it is essential that participants volunteer prior to random assignment. If the research is a comparison study about an implementation that is already underway, the selection of product group teachers need not have been voluntary, but the participation in data collection activities such as surveys and interviews should be.

If a teacher chooses to drop out of the study, it is appropriate for the researcher to understand the circumstances and motivation, but it is not appropriate to offer additional incentives or withhold other promised payments in the hopes of persuading the teacher to stay in.

Questions about Providing Incentives for Participation

It is common in effectiveness research to provide the product or service and related resources to the districts, schools, and teachers involved in a study to induce participation. Providing teachers or other participants a modest honorarium for their efforts involved with data collection, such as surveys required beyond regular classroom work, is also common. While teachers and other participants are commonly offered honoraria and other benefits, excessive inducements—especially if they favor the group using the product or service—may influence the results and should be avoided.

Districts are often interested in participating because a study can offer materials they otherwise could not afford. Such arrangements are appropriate, provided the benefit is not (perceived to be) dependent upon the study results. Excessive incentives may be considered a form of coercion.

It is important that the intervention and comparison groups are treated equally. For example, where teachers are the unit of analysis, each should receive the same honorarium for participating, regardless of group assignment. While it is often necessary to pay the cost of training time for teachers using an intervention, it is not appropriate to provide additional rewards for classroom implementation time. It is also appropriate to offer the intervention (and training) to the comparison group once the study is over.

Provision of free materials and generous incentives can sometimes be demotivating. Ideally, when the school is literally invested in the product or service, and therefore in the research, they are eager to find out whether a strong implementation will lead to the desired results. Research participants who are primarily motivated by monetary incentives or who do not have their own resources invested may be insufficiently concerned with fidelity of implementation (see Glossary) or with complying with the research requirements to carry out the research properly.

11 Pay attention to implementation, but not too much attention

It is well understood that poor implementation can make an otherwise effective product perform poorly. An explicit model of implementation (as recommended in Guideline 1) specifies the set of conditions under which the provider predicts that the product will have the greatest effect. In experimental evaluations, this is often called "fidelity of implementation." While it is important to ensure that product support systems are adequate to ensure fidelity, it is also important for the research to be designed around a sample that represents a typical and practical school system support and implementation pattern.

We have already talked in Guideline 4 about how part of planning for research depends on the stage of development the product is in, how well the training and support requirements are understood, and whether the goal of the research is to demonstrate efficacy or effectiveness. This Guideline is a reminder that in conducting the study, the product is being implemented (successfully or not).

Providing sufficient support for implementation is important because the full impact of an intervention may not occur in sites where the implementation model is not fully realized. Moreover, when implementation is consistent across all schools or classrooms using the product or service, the ability of an experiment to detect a difference will increase. Ideally, the provider works with the education customer to ensure that all the necessary support and training are provided and that the other conditions are met. In general, in the case of effectiveness evaluations, the researchers are not responsible for supporting the product or, in most cases, informing the provider when implementation is failing. Given an explicit logic model, the researcher can document both the support provided and the extent to which the implementation plan was followed.

Educational technology implementation occurs within very complex organizational structures of resources and people. Insufficient hardware access, too little time on task, lack of educator willingness and/or ability to appropriately integrate the technology, and inadequate school leadership and support can all negatively affect the implementation and the resulting impact.

With today's edtech products, which standardly collect usage and learning process data in cloud-based servers, it is possible to accurately measure whether a product is being used as intended and is functioning as designed. The researcher collects these data from the provider or from the district. The researcher can either implement additional support (in an efficacy study) or analyze successes and failures to assist with product and training improvements for the next iteration of the product (in an effectiveness study).

Researchers also have statistical techniques to take advantage of differences in implementation as detected by multiple usage metrics. The overall amount of usage can be analyzed as "dosage". Different metrics can be broken out to see what modes of usage produced the most positive (or negative) outcomes, which can then be fed back into improvements of the implementation model. In some cases, distinct patterns of usage can indicate different approaches to implementation providing insight into different impacts. Far from hiding product effectiveness from the researcher, differences in implementation can be used to better understand how and for whom the product works.

> With today's edtech products, which standardly collect usage and learning process data in cloud-based servers, it is possible to accurately measure whether a product is being used as intended and is functioning as designed.

12 Work with researchers who can be objective and independent

This Guideline is about maintaining credibility. This starts in the planning stage with the choice of researcher and includes the kind of contract put in place for the work, which determines the editorial and reporting process. These issues are present whether the research is conducted internally or through an external contractor.

Internal researchers or external contractors can conduct provider-sponsored evaluation research. In either case, to prevent undue influence (and the perception of it) and to help ensure objective and independent findings, explicit steps should be taken before work begins. The following are primary examples of such steps.

- Create a clear separation between the provider's internal research function aimed at producing publicly available evidence of effectiveness, and the provider's marketing and communication functions. However, formative product testing as part of continuous improvement need not be separate from marketing or product development.

- If the provider's internal research staff conducts the study, they will enhance credibility by reporting results and legitimately publishing their reports, regardless of the outcome. That credibility is even more enhanced if the report states that the researcher was given autonomy to publish before the study's data were collected.
- If an independent research organization conducts the study, they can enhance the credibility of their independence if they are given authorship and editorial control with a distribution license to the provider. Some providers, or their legal departments, may insist on a work made for hire contract with all rights retained by the provider (as though they were contracting for a piece of content). This is self-defeating since it will lack credibility as research.

Some edtech providers employ qualified researchers with expertise in research design and analysis, and who have received federal research grants. While research misconduct can be found even in prestigious institutions, researchers with direct commercial interests in the product or service may be more open to suspicion. The question is not necessarily about fraud, but instead about more subtle forms of bias, such as a tendency to emphasize results that are consistent with preexisting beliefs. Even outside contract researchers, whether working independently or conducting a work for hire, can be under suspicion if the next contract is perceived to depend on obtaining favorable results.

The best way to ensure high quality research is to work with researchers who know how to do quality work, are given independence to conduct the work, and who need to maintain a reputation for quality work.

An independently-funded outside research group provides the strongest assurance of objectivity. In this model, the provider assists an outside researcher with obtaining a grant from a government or foundation. Such funding, if following the ESSA levels of evidence, will require an existing body of independent studies. Unfortunately, this is seldom a viable option, as the supply of funding is limited relative to the demand driven by the many products and product studies.

The best way to ensure high quality research is to work with researchers who know how to do quality work, are given independence to conduct the work, and who need to maintain a reputation for quality work. A researcher, either internal or external, should be able to tell the provider whether the research is designed to meet a particular level of evidence established for ESSA or other purposes. For example, the researcher can say that the research is designed as an ESSA level 1 RCT. The provider must keep in mind, however, that the researcher may not have control of all the things that can go wrong in the implementation of the experiment (e.g., too much attrition or lack of student testing). And of course, the researcher cannot guarantee that the product will have an impact. A researcher with a track record of having conducted studies (not necessarily published in journals) indicates that both the design and operation of the experiment will go as planned.

IV.

Reporting the Results

The fourth set of guidelines focuses on good practices in reporting results from the research study. Research provides value for marketing, sales, and product development through the reporting of results. This reporting can take many forms from web-based animated presentations to publication in academic journals. The need for evidence on quickly evolving products and the availability of non-traditional repositories is changing the landscape of reporting. We also address the question of risk to the edtech provider in supporting research that shows a negative (or not sufficiently robust) effect of the product.

13 Produce reports with enough detail for decision makers to know if the results apply to their context

The report doesn't have to be a 50-page article with extensive background research and an appendix with many statistical data tables. However, certain essential elements in the report will serve as documentation of evidence at the ESSA levels. Other elements help school decision makers know if the results apply to their context. This guideline focuses on the most useful and practical reporting details for an audience of school decision makers and reviewers of grant proposals where a level of ESSA evidence is called for.

Elements of Practical Value

Reports will have practical value for school decision makers who want to know whether the evidence applies to their context.

Differential effects for subgroups. The main outcomes may consist of estimates of the average difference between the intervention and comparison groups. The results should also include estimates of the extent to which an intervention is differentially effective for different demographic categories or implementation conditions. For example, an overall result of no discernible difference between intervention and comparison groups (the main outcome) may mask a finding that the intervention worked very well for some part of the population but not for another (secondary outcomes). This can be essential information, both for educators and for providers, in understanding how best to implement the intervention and for the buyers to know whether the report is likely to apply to their schools.

What the product was compared to. The size of the impact reported (and even whether it was positive or negative) will depend on what the comparison group was doing. Sometimes researchers do not know this, but where possible, the report should detail the curriculum, program, and practices the comparison group used. While another educator implementing the product would not attempt to replicate the comparison group, it is important to know from what baseline the impact of the intervention was calculated. The report should provide key data for both the intervention and comparison groups, including the student demographics and average pretest scores so decision makers can recognize how similar the research setting and population is to their own.

Support for implementing the product. Since experimental evaluations can vary widely in their support for implementation, the report should clearly describe the product, the training and support delivered by the provider, the technology infrastructure, and additional support and resources provided by the school or district. The report should describe the amount of support, in comparison to a typical implementation, to enable readers to determine their ability to replicate the implementation.

Time frame and product version. The report should clearly state the time frame of the study. This applies to the school year of the implementation, the school year of data collection, and which version of the product was implemented (especially where new versions have important or large improvements). If a study conducted on an earlier version is cited to provide evidence of effectiveness for a later version or product, then the report should discuss the differences between the two.

"After the fact" exploration. Reporting explorations of the data that uncover additional outcomes can provide useful insight. Much can be learned by inspecting the patterns of results and identifying surprising relationships. Such examinations include finding whether a correlation exists between quality of implementation and outcome. Most importantly, analyses of this kind are considered exploratory, and firm conclusions should not be drawn from them. When exploratory findings are included, they should be labeled as such.

Whose report is it? A report should clearly state who initiated, funded, and exercised final editorial control over the research. For example, did the request for research come from a customer, potential customer, or provider? Did the provider initiate the study and recruit a school system to acquire generalizable evidence of effectiveness that can be presented to other potential customers? In other cases, a school district may have purchased the intervention and then conducted the research themselves or invited the provider or another entity to conduct research. Often, the provider will publish the report, therefore it is useful for the reader to know whether the reporting may be biased. Transparency in reporting is critical in demonstrating a study's credibility, particularly for reports published under the provider's masthead or otherwise viewed as controlled by the provider, as well as for those not attributed to a specific independent author.

Elements Needed for Technical Review

Conventional research reports include considerable detail to enable other researchers to replicate the original study or to combine studies of the same product or service into a research synthesis. For reports aimed at a non-technical audience, these details, if included, can be assigned to an appendix or included in a complete version that is referred to in a shorter, more user-friendly report of the findings. Here we outline the requirements for a report to be reviewed for compliance using ESSA evidence standards. The outline is only meant to suggest the kinds of information called for, not to provide technical instructions to researchers who will need to refer to documents from organizations such as the WWC, which is specifically named in the regulatory guidance.

We address the requirements for reports to document evidence at each of the ESSA levels.

For ESSA level 4: Base level. This level does not call for field research but it is an essential preliminary to subsequent research. The logic model (described in Guideline 1) is a convenient way to organize and record the rationale based on research on learning and on similar products. Many grant programs for which school customers apply require this level of rationale for products where there is not yet any field studies.

For ESSA level 3: Promising. The primary requirement to show promising evidence of impact is a correlational study, described in Guideline 7. The research report must explain which important element of the logic model (which can be as simple as "sufficient usage") is to be correlated with what outcome (also mentioned in the logic model). There are fewer strict rules for reporting such studies, but one requirement is for the variables to statistically control for what ESSA calls selection bias. A study may report the correlation of an important aspect of usage with the outcome if the statistical controls for factors that could account for selection into the user group are in place.

For ESSA level 2: Moderate. WWC specifies the requirements for these reports. They must include: (1) a table to demonstrate that the two groups being compared are closely matched at "baseline" (within 0.25 of a standard deviation), and (2) the pretest of the measure being used as the outcome. Selection bias is also an issue here, so the table showing how closely the groups are matched should include factors that plausibly account for selection.

For ESSA level 1: Strong. The requirements for these reports are specified by the WWC and are quite detailed, including the portion of units that can be lost (attrition), which, if occurring more in one condition than the other, may be a sign that the study could be biased. For example, if a product or service results in more low-scoring students dropping out of the intervention group than the comparison group, the outcome may be affected by the attrition. Many other technical elements not required for levels 2 or 3 are expected to be included in this report.

For any report at levels 1-3. The report must be clear about the limitations of the study, especially with respect to generalizability. All studies have limitations, and clearly presenting these limitations is necessary for readers to understand how to use the research for decisions within their local context. The limits to generalization must be stated in relation to the following: a full description of the sample, the comparison group, student characteristics, teacher characteristics, and setting.

When marketing materials refer to specific research findings or to product impacts based on evaluation research, they should link to the full report so that the reader can put the findings in context and directly review and evaluate the claims being made. Taking results out of the context in which they were observed can imply a greater generalizability than is warranted by the original study. For example, a graph may be taken from a report to illustrate an effect, but without a reference back to the context of the research finding, the graph may be misleading. The original report should always be freely accessible.

14 Make the research report easily accessible and invite external review

These guidelines take the position that traditional peer-reviewed journals are generally inappropriate for product effectiveness research, although they continue to be important for scientific research, including advances in methodologies conducted in higher education institutions and similar organizations. As the pace of research (not just research on edtech impact) accelerates, the need for more open resources and repositories increases. There are a growing number of alternatives to traditional journals.

Making Research on Edtech Impact Accessible

There are three reasons that scientific journals may not be the optimal or most accessible venue.

1. Using scientific methods to show a product's effectiveness is not itself a contribution to the kind of science in which journals are usually interested. Effectiveness research does get reported in them, but this is when there is a unique aspect to the product or the methodology used.

2. The review and revision process, i.e., peer-review, is usually slow and time consuming for the authors. While conferring prestige to the report, it does not in most cases guarantee acceptance as evidence under the criteria spelled out in ESSA.
3. Journals generally require a subscription, making them inaccessible to company personnel or educators outside of academic institutions. There are growing exceptions to this picture, with "open" journals and the requirement by some foundations and federal agencies that research they fund be made freely available.

The academic journal still retains a prestige within the university research community, and while members of that community are on the review panels of foundations and government agencies that award funding for education projects these journals will continue to hold weight. However, alternatives are appearing for easier accessibility and for a quality review that may better suit the volume and speed of research on edtech impact.

Here are some common approaches.

Publishing a report on the provider's web site is the most common way to make the research results accessible. This may have the disadvantage of appearing biased since it is hosted on the same site as the product promotion. When research is conducted by an independent researcher, the credibility will strengthen if the report (or a full version of the report) is available on the researcher's website or other accessible repository used by the researcher. Repositories have emerged that provide web publishing. Two examples are SSRN and Academia, both of which provide free posting and access. Both allow visitors to view other reports by the same author. Academia includes advertising and elements of social media connecting authors to readers. Neither provide ratings based on formal reviews, although they provide rankings of numbers of downloads and other popularity metrics. A long-standing but simpler version of a repository is supported by ED: Education Resources Information Center (ERIC). Reports posted on repository sites are not considered published in the way that would prevent them from later appearing in journals as original articles. Unless copyright restrictions prevent doing so (e.g., where the research has been published in a scientific journal), the report should be provided for free download.

Many conferences provide a professional audience for research results, and in many cases, provide a repository of papers presented at the conference. Some examples include American Education Research Association, American Evaluation Association, and Society for Research on Educational Effectiveness. Acceptance to present at the conference offers a version of peer review and recognition as a contribution to a corpus of research curated by these organizations.

Getting Research Reviewed

Peer review for scientific journals is common in the academic community. It allows for critique by others and provides the opportunity for revisions and clarifications to ensure that a study meets research standards in terms of its methods, claims, and so forth. At the end of such a review process, a study should be worthy of acceptance into the corpus of scientific work. Journals, however, primarily publish theoretically or methodologically oriented research, where the question being addressed arises from a researcher's career-long program of research. This process is generally too cumbersome to be useful for industry.

A version of peer review has also emerged in the form of government-funded organizations such as the WWC, which focus on educational programs, including edtech products. Their focus is on evaluating the likely effectiveness of programs through review of research. They are not a repository for research. Unlike academic journals, these organizations actively seek research reports applicable to the domains in which they are conducting reviews. Also unlike academic journals, their review is more formulaic. Rather than engaging in a back-and-forth process of modification, as with a traditional research journal, these initiatives simply review and rate the study depending upon the degree to which it meets their explicit guidelines for acceptable research.

Crowdsourcing reviews, which could be a solution to getting prompt, accurate reviews, may become available. Posting a report online and requesting comments may be a simple approach. Education magazines and newsletters may become discerning about research quality. The broad dissemination of these Guidelines may encourage more careful examination of reports, raising questions that researchers may need to respond to in order to maintain credibility.

15 Make all findings from product evaluations available, as a general rule

Conducting a rigorous study of product effectiveness can be a serious risk for an edtech provider. If the experiment shows a positive effect, that is good news, but if there's no discernable effect, or worse, the comparison group does significantly better, this may be a blow to a planned marketing campaign. In the fast-paced world of edtech development, there are approaches to mitigating this risk, most importantly, not depending on a single study by setting up a program of research. But it is important to understand that not making a study public can reduce educators' trust in the provider's product and in industry research as a whole.

Following from the misguided belief that good experiments are very expensive and take a long time, the conventional approach to research assumes a single study has given a thumbs-up or down on the product. With the acceleration of edtech development and continuous improvement of products, providers have to start thinking in terms of a program of research where each study, whatever the ESSA level, reflects back on improvements and provides buyers with evidence as to how and for whom the product works best.

Why, Generally, All Results Should Be Reported

An important goal of sharing research is to build trust between the provider community and consumers of education products. Scientific research increases knowledge over time using multiple replications of experiments to test hypotheses under a range of conditions. In a program of research that encompasses a variety of methods, populations, and studies, not all results will be positive, and not all sites will be able to implement the intervention with fidelity. Still, it is hoped that the preponderance of evidence should demonstrate the intervention's impact. Reporting results that are less positive will help the stakeholder community, including researchers and educators, to attribute greater credibility to research efforts as a whole.

Thus, an ideal in scientific practice is that research findings are made available regardless of the result. We can see how a company conducting a study in all 50 states, and only publishing those that showed an impact, could produce bias. Even if the product has no effect, two or three of the studies are likely to show significant differences just by chance. While this is an extreme example, it is a common problem, whether in education or any other field of scientific research. Researchers and research journals tend to prefer reporting studies with positive results. In medicine, where failed trials are under reported, a product may appear more effective or less dangerous than it really is. This is a well-known publication bias. In the world of edtech providers, where there is far less regulation of experiments than in medicine, the underreporting of research can strain the confidence of consumers in the results. In cases where consumers find out that results have been suppressed, confidence in the published results will be diminished.

This Guideline states that edtech providers should follow a general rule: to publish all research. The Society for Research on Educational Effectiveness (SREE) is establishing a registry of research, so that researchers who register their study in advance, as is done in medicine, will get credit or credibility for not hiding negative findings, that is, for taking the risk that the findings may not show positive results. Reporting results is not mandatory but may raise questions about whether there was a valid exception.

Exception for Formative Research

There are circumstances where it makes sense to not publish the result. For edtech products that are continuously being improved, studies should not just show if the product had an overall impact, but whether the impact was different for different students, teachers, or schools. It should help the provider understand how the product had an effect, including the strengths and weaknesses of the product. In cases where the study has value for product improvement, and where the provider intends to make improvements and try again, the original report will have little value for customers of the improved version. Following this logic, there's no problem conducting a series of formative experiments, improving the product at each stage until positive results are found.

This is very different from conducting a series of experiments on the same version of the product until positive results are found. Without improvements, the research process is just fishing for results that may be a random result. Improvements made by providers may be in the software or in the support, training, recommendations, or implementation, or also by targeting to populations where the product works.

Exception for a Failed Experiment

Conducting research in schools is difficult, given the many practical challenges, which may simply include implementing the product or service. The following cases consider an evaluation to have failed, and therefore, results are not reported:

- Implementation of the intervention clearly failed, meaning that it was implemented with such low fidelity that it would be unfair, inappropriate, and misleading to report these results;
- A critical piece of the planned data collection was blocked, such that results could not be determined;
- Or the sample was insufficient in size or biased, resulting from an inability to identify or gain participation or from severe attrition among study participants.

Legitimacy for aborting a study and not reporting it under these conditions requires that this determination be made before the outcome measures are collected and inspected. Otherwise, withholding the study's report may be perceived as driven by poor results, rather than by the decision that the experiment failed. In most cases, flaws in the experiment should not be grounds for withholding the results, although appropriate disclaimers should be made.

Exceptions to the Guideline that providers should report results regardless of the outcome occur in cases of product improvements and failed experiments. However, holding reports back on versions of a product or service currently in the market or on which results are reported elsewhere, simply because of unfavorable results, is not among the exceptions.

16 In the marketing literature for a product, accurately describe its impact in non-technical language

This Guideline addresses the translation and communication of research findings to other parties. Although educators with little formal training in research methods may be the primary audience, many school systems have staff trained in research methods who will want to compare an intervention's marketing claims to what is found in the full report. It is important that reporting the research does not overstate what has been established through rigorous analyses. Otherwise, research and marketing claims will lose credibility over time.

It is important to translate research findings into language that educators without an advanced degree in research methodology can understand. At the same time, it is essential that some of the complexity and conditionality of the results be communicated. Provider staff who are responsible for customer communication may find translating formal research into understandable and appropriate product claims to be a challenge. Tools are not readily available to assist them in making complex research findings clear to potential customers within the time they have for explaining them.

Providers should address these issues in the early stages, for example, by including in the contract that they also receive a non-technical summary of the results.

If the provider's internal research staff lacks the qualifications, then the external researchers employed by the provider may be willing to assist in this task. However, when external researchers have completed the study, they may not be available or willing to assist the provider with explaining the results or by reviewing the accuracy of the provider's translation. Submission of the full report of the study may be the last of their contracted responsibilities. Moreover, they may consider helping develop marketing materials to be in conflict with their commitment to objectivity. Providers should address these issues in the early stages, for example, by including in the contract that they also receive a non-technical summary of the results.

It is useful to align marketing literature with ESSA levels of evidence, so that claims of causation are substantiated by appropriate designs that can eliminate plausible alternative explanations for the findings. Where doubts remain about the extent to which the research eliminated other plausible explanations for observed achievement gains, it is preferable to suggest that a strong association exists between introducing the product or service and the observed gains. "The study found that our product was associated with higher achievement levels" reports a correlation (ESSA level 3), whereas "The study found that our product had a significant impact on achievement levels" makes a stronger causal statement, calling for a level 1 or 2 study.

Claims referencing the strength of the impact should use language that reflects an effect's size and its educational meaningfulness. Where appropriate, evaluation findings should be translated into terms of practical significance (e.g., test score percentiles or dollars per student). Researchers usually report impacts in terms of "standardized effect sizes," and it is important that these be included in the full report. Translations into percentile rank changes are straightforward.

CONCLUSION

In revising the Guidelines, originally released in 2011, we are aware that the technology being used in edtech has changed and continues to change at an accelerating rate. We hope that the Guidelines will help providers understand research as an ongoing process, rather than as a one-time activity. This is especially important considering the speed of technology innovation and new product development, which will often outpace the research cycle and educators' calls for evidence of effectiveness. Traditional tools for the review and dissemination of research are generally not able to keep up with new versions of products or services. Thus, we have tried to paint a picture of continuous improvement, where research reports have a short shelf life, and the research process can provide both solid evidence of effectiveness, and at the same time, provide feedback for product improvement.

Further changes to technology and research processes are inevitable. The K-12 marketplace consists of thousands of edtech products, and the schools are faced with balancing the value of many alternative ways to support student learning, teacher development, and administrative functions. Conventional research approaches are not adequate for the thousands of studies needed to keep up with inventory or continuously improving products. The traditional academic review and publishing process will not be timely and relevant, and new kinds of research repositories will be needed. SIIA plans to continue tracking changes in research practice and providing guidance to its members and the industry as a whole. Our hope is that these Guidelines will help providers understand the current standards of research practice. As these standards continue to evolve, we can continue to track evidence requirements that educators and developers will find both productive and workable.

GLOSSARY & RESOURCES

Aggregate data

In research on schools, aggregate data refers to data on student test scores, demographics, or any other measurable characteristic summarized at the classroom, grade, or school level. One advantage of aggregate data is that it eliminates personally identifiable data, therefore it can be reported publicly. The downside is that analyses using aggregate data are less sensitive: estimated effect sizes are lower than those estimated at the student level by the factor of two or three. In addition, data analysts must use caution in interpreting results since establishing an association between an impact and an aggregate characteristic—for example, finding that the impact is stronger in schools with higher percentages of English learners—does not necessarily mean that the impact on individual students (English learners) is stronger compared to other students within schools.

Confounding factors

The goal of an impact study is to identify factors that directly affect outcomes. Researchers have developed study design and analysis techniques that help disentangle the effects of various factors influencing individual learning and school environment. There is no remedy, however, if all treatment units—that is, all students in schools or classes using the product or program under investigation—have the same characteristic, or are affected by some change that does not affect the comparison units. If this is the case, it is impossible to distinguish between the treatment effect and the potential effects of other confounding factors. When designing a study, the researcher must carefully consider all the variables that could affect their findings, otherwise the results of their study might not be valid. For example, let's say that a developer decided to test out a Positive Behavior program within a district that contained 4 middle schools. Two schools used the program throughout the school year, and the other two did not. At the end of the year, the two schools that used the program reported fewer discipline referrals than the other two. Based on the results, one might attribute the difference in discipline referral rates to the program. However, it turns out that due to a new state legislation that changed certain funding formulas, the two schools received additional funding that enabled reducing class sizes and hiring additional support personnel. The change in funding is a confounding factor: an unexpected and unaccounted-for change that affects several important variables and damages the internal validity of the study.

Correlational designs and regression methods

Correlational studies can determine whether there is a statistical association between two variables (and how strong this association is) but cannot establish causality. For example, a correlational study may find that there is a strong positive relationship between time spent on an online math program and test scores (the more time students spend on the program, the higher their test scores), but one cannot be sure that the higher test scores are directly caused by spending more time on the program. Somewhat counterintuitively, correlational studies do not usually rely on statistical analysis of correlation, but instead use methods of regression analysis. This makes it possible to adjust the association between the outcome and the metric of interest (such as time spent using a program) for influences of various variables relevant to learning. Thus, a regression model would include, in addition to the variable of interest, the time spent on the product, and factors such as pretest score, gender, English language fluency, etc. Unlike experimental studies that produce results showing actual differences between treatment and control groups, correlational studies produce results that have a hypothetical flare, stating by how much the outcome changes in association with a one-unit change in the variable of interest, for example, the average test score difference associated with a one-hour difference in time spent on the product. This presentation of results does not imply that increasing the time by one hour would cause the given increase in the test scores, or that a one-hour increase is even feasible. Results of correlational studies can therefore only show promise of effectiveness, not prove the impact. However, regression analysis may be considered marginally stronger than a basic correlational analysis at indicating causal links. Specifically, it allows a basic adjustment or correction for the effects of variables that are confounded with the treatment and that influence outcomes. Such adjustments, however, give no assurance that all relevant confounds are accounted for, and that the result, with adjustment, is unbiased or less biased compared to an unadjusted result. Empirical work has shown that the quality of covariates matters for reducing bias. It is important to adjust for the pre-intervention measure of the outcome (i.e., the pretest) and other variables that are indicative of individuals' motivations for selecting into the program.

Effect size

This is a common way for researchers to express the strength of impact in a way that does not depend on the units of measurement and, therefore, can be easily compared across studies. It is defined as the difference in means of the outcome variable between the treatment and control groups—the estimated impact—divided by the pooled standard deviation of the outcome variable. The standard deviation is a measure of variability in the data (see entry later). The effect size thus helps to understand by how much the treated group moves up in relation to the distribution of outcomes due to the treatment. Researchers often use notions of small, moderate, and high effect sizes, but no consensus exists. Effect sizes under 0.10 are commonly referred to as small: only the most rigorous experiments can detect effect sizes as small as 0.05. It is very unusual to see an impact as large as 0.5 (half of a standard deviation) in field studies, although not uncommon in laboratory studies. Laboratory studies are more likely to test the treatment against no treatment, similar to a medical placebo (see Guideline 2's discussion of Comparison Groups). The ESSA definition mentions "a difference of 0.25

standard deviations or larger" as the threshold of "substantively important." The 0.25 threshold is arbitrary, and the ESSA definition fails to distinguish laboratory studies from field studies and puts no requirement for statistical significance if the result passes the 0.25 threshold. No matter what the effect size is, the logic of statistical analysis suggests that only statistically significant results are worth considering. We have been told by those familiar with the drafting of the WWC rules, on which the ESSA definition is based, that they presumed that an experiment that demonstrated substantively important effect would probably have a credible level of significance.

Efficacy vs. Effectiveness

As described in Guideline 4, efficacy studies show how a product or service works in the *ideal case*. In an efficacy study, the goal is often product improvement rather than, or in addition to, initial measurement of impact. By contrast, an effectiveness study provides no more than the usual level of training and support that an ordinary customer would receive, as they reflect *ordinary conditions* of implementation.

In 2013, the ED's Institute of Education Sciences (IES) and the National Science Foundation (NSF) developed "Common Guidelines for Education Research and Development." They describe six types of research, and define efficacy and effectiveness research as the following:

> "*Efficacy Research* allows for testing of a strategy or intervention under 'ideal' circumstances, including with a higher level of support or developer involvement than would be the case under normal circumstances. Efficacy Research studies may choose to limit the investigation to a single population of interest… *Effectiveness Research* examines effectiveness of a strategy or intervention under circumstances that would typically prevail in the target context. The importance of 'typical' circumstances means that there should not be more substantial developer support than in normal implementation, and there should not be substantial involvement in the evaluation of the strategy or intervention."

Fidelity of implementation

Fidelity of implementation in a school system setting means accurate and consistent delivery of content and instructional strategies in the way in which they were designed and intended to be delivered by the program or product developer. As discussed in Guideline 11, it is important to ensure that product support systems are adequate to ensure fidelity of implementation, however it is also important for the research to be designed around a sample that represents a typical and practical school system support and implementation pattern.

Formative research

Formative research takes place during the product development cycle in order to garner feedback to better influence design decisions, as well as to guide the product formation efforts. With formative research, user input and suggestions can be incorporated before the product is released to a broader audience. This type of research does not qualify as "evidence-based" under the definitions provided by ESSA, but nonetheless plays an important role for edtech developers as they continuously improve and refine their technology.

Impact research

As described in IES and NSF's Common Guidelines for Education Research and Development, the purpose of Impact Research is to generate reliable estimates of the ability of a fully developed intervention or strategy to achieve its intended outcomes. For an impact study to be warranted, the theory of action must be well established, and the components of the intervention or strategy well specified. The three types of impact studies—Efficacy, Effectiveness, and Scale-up—differ with regard to the conditions under which the intervention is implemented and the populations to which the findings generalize. In addition, as the research moves from Efficacy to Scale-up, studies should also give greater attention to identifying variation among impacts by subgroup, setting, level of implementation, and other moderators. For all impact studies, descriptive and exploratory analyses should be sufficiently elaborated on to determine the extent to which the findings support the underlying theory of action.

Institutional Review Board (IRB)

Many research organizations, universities, and larger school districts have a committee that reviews research. IRB contractors can provide the review service to smaller companies. The IRB's role is to assure that people who may be participants—including students and teachers—are adequately protected and that, where necessary, informed consent is obtained. The IRB decides whether parental consent is needed. Where the research involves ordinary classroom or school activities, parental consent is often not required. Where the research uses only data that have already been collected ("extant data")—such as school district administrative data and test results—and where the data are deidentified—or where accidental disclosure represents minimal risk—a review may not be called for or only an expedited review will be needed. The IRB addresses risks to the people involved. This is separate from additional laws, regulations, and policies that govern disclosure of student data.

Intent-to-treat (ITT)

Intention-to-treat (ITT) analysis of the results of an experiment is based on the initial treatment assignment, and ignores the discrepancies between the plan and the treatment actually received. ITT analysis is intended to avoid pitfalls that can result from problems with study implementation, such as non-random attrition of participants. ITT is also the simplest approach, because it does not require either collecting data on compliance of study units to the initial design or making adjustments whenever non-compliance is detected. For example, a teacher whose classroom is assigned to the treatment group may decide that he or she does not have the bandwidth to participate and does not end up using the product. His or her students are still considered part of the treatment group in the final "intent to treat" analysis because of their original assignment.

Laboratory research

Much of the developmental or educational research conducted by universities examines learning or cognitive processes using a relatively small number of students and in carefully controlled conditions. Such studies can be classified as learning science, and many reviewers do not accept such studies as evidence of impact of products because it is very easy to obtain very large effect sizes and meet the "substantively important" criteria while not showing whether the product works in the field. For this reason, some reviewers will not accept studies shorter than 12 weeks from program inception to posttest, or studies with fewer than 120 students. ESSA is more stringent, requiring a sample of at least 350 students for a study to meet the standards of level 1 or 2.

Logic model or Theory of action

The ED's Electronic Code of Federal Regulations (e-CFR) describes a logic model—also referred to as theory of action—as "a well-specified conceptual framework that identifies key components of the proposed process, product, strategy, or practice (i.e., the active 'ingredients' that are hypothesized to be critical to achieving the relevant outcomes) and describes the relationships among the key components and outcomes, theoretically and operationally." The W.K. Kellogg Foundation offers a helpful resource for developing logic models. Intended mainly for non-profit organizations undergoing project evaluation, the document nonetheless provides a comprehensive overview and instructional guide for developing a logic model, including templates and checklists.

Personally identifiable information (PII)

The FERPA definition of personally identifiable information (34 CFR § 99.3) follows the government-wide definition.

Personally identifiable information includes, but is not limited to:

4. The student's name;
5. The name of the student's parent or other family members;
6. The address of the student or student's family;
7. A personal identifier, such as the student's Social Security Number, student number, or biometric record;
8. Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
9. Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty; and
10. Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.

Pilot testing

School systems will often pilot an edtech product to test out the technology within their setting and gather teacher and student feedback before committing to a school-wide or district-wide purchase. Digital Promise, an independent nonprofit specializing in innovative education practices, offers a resource called the "Ed-Tech Pilot Framework," which is especially helpful for school-system leaders and other administrators who are in charge of purchasing decisions. The framework provides an 8-step process to help both education leaders and technology developers run successful educational edtech pilots. The framework is supplemented with case studies, research reports, and other resources.

Quasi-experiment

The ED, through the Electronic Code of Federal Regulations (e-CFR), describes a quasi-experiment (QE) as “a study using a design that attempts to approximate an experimental design by identifying a comparison group that is similar to the treatment group in important respects. These studies, depending on design and implementation, can meet What Works Clearinghouse Evidence Standards with reservations (but not What Works Clearinghouse Evidence Standards without reservations).” The Institute of Education Science published a helpful report that describes this technical process called “Statistical Power Analysis in Education Research.” In this paper, researchers Hedges and Rhoades define QE as: “A research design that compares groups but does not involve randomization. Rather the treatment and comparison groups are often matched based on pretests or demographic factors, such as socioeconomic status.”

Randomized control trial or Randomized experiment

The highest ESSA level of evidence—level 1: strong evidence—involves random assignment of units (e.g., teachers, schools) to using the product or not. The ED, through the Electronic Code of Federal Regulations (e-CFR), describes a randomized controlled trial (RCT) as “a study that employs random assignment of, for example, students, teachers, classrooms, schools, or districts to receive the intervention being evaluated (the treatment group) or not to receive the intervention (the control group). The estimated effectiveness of the intervention is the difference between the average outcomes for the treatment group and for the control group. These studies, depending on design and implementation, can meet What Works Clearinghouse Evidence Standards without reservations.”

Regression analysis

Regression analysis is a group of analytical methods applied to situations, in which outcomes can be influenced by multiple interrelated variables. The result of this analysis provides estimates of the strength of association between each variable (characteristics of study subjects and product usage) and the outcome, holding other things equal. The use of regression analysis is particularly important in non-experimental studies where the goal is to estimate the effect of product usage on the outcome but where any number of things can drive both the product usage and the outcome. A regression analysis allows controlling for the intervening influences by “taking them out of the equation.” Controlling for pretest—i.e. variability of the students’ level of preparation at the outset of the study—is particularly important. The use of regression analysis involves specifying a model—a statistical equation that relates the outcome to the variables of interest and other influences that are summarily called covariates in this context. In addition, researchers make assumptions about the error term, that is, the possible ways in which unmeasured factors could affect the results. In the studies of educational effectiveness, hierarchical linear models are used. These models assume a linear relationship between the covariate and the outcome and take into account the following: that students are grouped into classes; that schools are grouped into districts; and that units within one group tend to be more similar to each other than to units in other groups and are therefore influenced by the same factors of which we are unaware.

Selection bias

Selection bias is caused by non-random assignment of study subjects to the treatment group. In other words, it results when students, teachers, or schools decide themselves whether they want to join a non-experimental study, or drop out of a controlled experiment after the initial random assignment to treatment and control groups is made. Such voluntary decisions are often determined by certain characteristics of study subjects and therefore lead to a situation where the apparent positive study impact is in fact a result of the treatment group being composed of better prepared or more motivated subjects. Hence, the "bias" in the impact estimate. Selection bias normally arises in the process of edtech adoption, as most such products allow a lot of freedom to their users. Consider a case where students who "selected" to read a lot of stories presented in a product did well on the reading test. The decision to read stories and the ability to do well on the test may have both been a result of previous enrollment in more advanced classes. If the researchers could not measure and control for that characteristic, they would come to an incorrect conclusion about the effectiveness of this product. Quasi-experimental study designs and bias-correction analytical methods have been designed to minimize the adverse effect of selection bias on the validity of the results. However, without randomization, we can never be sure we have measured the factors that are really determining selection. Selection bias is used, without definition, in the ESSA law where, for example, a requirement for level 3 evidence is a "well designed and well implemented correlational study with statistical controls for selection bias." Without further definition, which is not provided, this has been interpreted as requiring controlling at least for the pretest of the measure used as the outcome.

Standard deviation

Standard deviation is a basic concept in statistics. It is a measure of variability in the data or the extent of dispersion of the observed values around their mean; for example, student test scores can be widely distributed or more bunched together. Standard deviation is the basis for calculating effect size and determining statistical significance, as well as many other statistical measures. Every result obtained in effectiveness research should be reported together with its standard deviation or other measures that are based on it.

Statistical significance

Statistical significance is a measure of the likelihood of getting a result with a magnitude as large as (or larger than) the one observed in the study merely by chance, if there really was no difference or if the correlation really was zero. In other words, it is a measure of credibility of a study result: the higher the significance level, the lower its credibility, and the higher the probability of mistaking a random fluke for a real impact. This measure can and should be obtained for any estimate resulting from statistical data analysis since there is always some uncertainty in any result based on real-world data. The usual standard for statistical significance is 5% or less; there is less than a 5% chance (often expressed as $p < .05$) that the size of an impact was random variation or was the result of something not measured in the study. This strict standard is accepted in most government research, including that sponsored by ED. The idea is to be really confident that the product works before recommending it. The level of confidence required for a finding to count as evidence is not defined in ESSA or other regulations. As the basis for a decision by an educator to purchase a product or not, the 5% standard may lead to rejecting products that work, but that the researcher can't be highly confident about. So, often in studies of product impact, when results are intended to influence practical decisions, they are reported when there is a significance up to 20% as a conclusion with low confidence.