

***Conducting and Reporting  
Product Evaluation Research:  
Guidelines and Considerations for  
Educational Technology  
Publishers and Developers***

**Author: Denis Newman, Empirical Education Inc.  
Producer: Mark Schneiderman, SIIA**



**A Publication of the  
Software & Information  
Industry Association (SIIA)**

## **About SIIA**

The Software & Information Industry Association (SIIA) is the principal trade association for the software and digital content industries. SIIA provides global services in government relations, business development, corporate education and intellectual property protection to 500 leading software and information companies. SIIA's Education Division serves and represents more than 150 member companies that provide software, digital content and other technologies that address educational needs. The Division shapes and supports the industry by providing leadership, advocacy, business development opportunities and critical market information. SIIA provides a neutral business forum for its members to understand business models, technological advancements, market trends and best practices. With the leadership of the Division Board and collaborative efforts with educators and other stakeholders, the Division undertakes initiatives to enhance the use of educational technology and the success of SIIA members.

Software & Information Industry Association  
1090 Vermont Ave, NW, 6<sup>th</sup> Floor  
Washington, DC 20005-4095  
Phone: 202-289-7442  
Fax: 202-289-7097  
<http://http://www.sii.net>

## **About the Author**

Denis Newman founded Empirical Education, a seven-year-old educational research firm in Palo Alto, California, with the mission of helping school districts improve their decision-making. With a 35-year career in education, Newman draws on his awareness of children's learning in the classroom, extensive experience in developing instructional technologies and lifelong commitment to scientific research. Newman's Ph.D. in Developmental Psychology is from the City University of New York. He has conducted research and development at Rockefeller University, University of California – San Diego, Bank Street College of Education and BBN Corporation. He has published more than 35 books and articles and has served as program chair for the American Educational Research Association's Curriculum and Learning Division. He was a pioneer in the application of Internet technologies for student learning, professional development and school administration, bringing to market the first integrated web server designed for schools. His business career has included senior positions at educational software companies Tegrity and Soliloquy Learning.

## **Acknowledgements**

SIIA wishes to extend its thanks and appreciation to a number of individuals for their contribution to these Guidelines.

We extend our gratitude to our author, Denis Newman (Empirical Education), for his knowledge, commitment to excellence and persistence in the face of many rounds of discussion and review.

Again, our thanks to Denis and to his co-chair of the SIIA Education Division's Research & Evaluation working group, Rob Foshay (Texas Instruments), for their vision for these Guidelines and their leadership and tireless efforts in guiding them through to completion.

We also appreciate these members of the SIIA Education Division's Research & Evaluation working group for their support of this initiative, and for their ongoing input and review:

- Mick Adkisson, SMART Technologies
- Christopher Brown, Pearson Education
- Andrew R. Coulson, MIND Research Institute
- Valerie Crawford, SRI International
- Kristin DeVivo, Scholastic
- David Dockterman, Tom Snyder Productions
- Katie Gilligan, TextHelp
- Carol Guziak, Houghton Mifflin Harcourt
- Dana Jeffrey Laursen , PLATO Learning
- Irene McAfee, McGraw-Hill Education
- Steve Ritter, Carnegie Learning
- Sara Sawtelle, Learning Enhancement Corp.
- Jay Sivin-Kachala, IESD
- Cathy Sump, Curriculum Advantage

Finally, for their thoughtful comments on earlier drafts of this document, we wish to thank:

- Larry Hedges, Board of Trustees Professor of Statistics and Social Policy, Institute for Policy Research, Northwestern University;
- Robert Slavin, Co-Director of the Center for Research on the Education of Students Placed at Risk, Johns Hopkins University; and
- Talbot Bielefeldt, Senior Research Associate, International Society for Technology in Education.

SIIA and the primary author take responsibility for all the errors that remain.

# Conducting and Reporting Product Evaluation Research: Guidelines and Considerations for Educational Technology Publishers and Developers

## Contents

<i>About SIIA – About the Author -- Acknowledgements</i> .....	1
<i>Introduction</i> .....	4
<b>Purpose and Audience</b> .....	5
<b>Scope and Limitations</b> .....	5
<b>Challenges Common to Research on Educational Technology</b> .....	6
<i>Summary and Outline</i> .....	10
<i>Guidelines and Considerations for Conducting and Reporting Product Evaluation Research</i> ....	13
<b>Ask the Right Question</b> .....	13
<b>Support the Implementation of the Product or Service</b> .....	15
<b>Plan a Study of Sufficient Size and Duration to Demonstrate an Effect</b> .....	17
<b>Plan for Plausible Causal Claims</b> .....	21
<b>Avoid (the Appearance of) Conflicts of Interest</b> .....	24
<b>Provide a Comprehensive and Detailed Research Report</b> .....	27
<b>Make the Research Findings Widely Available</b> .....	31
<b>Accurately Translate Research for Customers</b> .....	34
<i>Conclusions</i> .....	36

# Introduction

*Conducting and Reporting Product Evaluation Research: Guidelines and Considerations for Educational Technology Publishers and Developers* (i.e., the Guidelines) grew out of a project of the SIIA Education Division's Research & Evaluation Working Group. The Guidelines seek to provide a standard of best practices for conducting and reporting evaluation studies of educational technologies in order to enhance their quality and credibility and, therefore, their utility to education decision makers. The focus is on publisher-developer sponsored or supported studies, but many guidelines apply more broadly. The brevity of this document relative to the complexities of research suggests that the Guidelines should not be considered to impose any absolute or binding requirements on research, but should instead be viewed as a general primer of best practices.

A number of factors, including the federal emphasis both on accountability and on research have combined to increase the demand by education decision makers for research evidence about their programs, practices, products and services. This includes research on the impact of the educational software, digital content and related technology or technology-based products and services, which we will refer to collectively as "products and services" or "interventions," provided by SIIA member publishers and developers. This trend is expected to continue among educators, whatever the evolving public policy requirements.

SIIA members and other product or service providers have responded by enhancing the scale, scope and rigor of their existing research

investments, including further documenting the scientific basis of their products and services and commissioning additional evaluation research. In so doing, the industry has come to recognize the value of establishing a set of research guidelines. This document is produced with guidance from SIIA's working group and in collaboration with other stakeholders and experts. The document also aims to help inform all stakeholders of the research challenges unique both to studies of technology and to vendor-commissioned research in general.

These Guidelines are based on an understanding that this evaluation research must be:

- relevant to educators by providing information needed to inform their decisions;
- transparent and complete in its disclosure and reporting;
- free of undue bias; and
- conducted using accepted research methods.

The Guidelines are not comprehensive in scale or scope, and they do not constitute a step-by-step manual for designing and conducting evaluation research studies. Instead, they attempt to flesh out the four basic ideas listed above. Moreover, although these Guidelines are geared to K-12 education and technology-based products and services, the principles they embody apply to other education sectors as well as, in many cases, to non-technology products and services.

## ***Purpose and Audience***

The Guidelines have several purposes and audiences. First, the Guidelines are intended primarily for publishers and developers (i.e., providers) of educational technologies, including especially software applications, digital content, e-learning and related instructional technology products and services. In particular, the Guidelines are written for the managers responsible for development, evaluation and marketing of these products and services, whether or not they have evaluation researchers on staff. The Guidelines address operational decisions – planning, designing, conducting and reporting – that are under the control of providers carrying out or commissioning a study on their products and services.

To provide information of greatest value, the Guidelines focus on issues that are most unique to educational technology products and services, and are not particularly well addressed in the general literature on evaluation design and methodology. These include determination of appropriate outcome

measures, the importance of fidelity of implementation and issues that are specific to vendor-sponsored research.

In addition, we hope the Guidelines will also give 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 technology-based products and services. Research reports that adhere to these Guidelines will be of high quality and credibility.

Finally, we expect the Guidelines to be reviewed by additional stakeholders – researchers, policy makers and education officials. For all audiences, the Guidelines not only provide a set of standards of practice, but also seek to advance the field by helping to identify the appropriate balance between the rigor, practicality and usefulness of evaluation studies of technology-based products and services.

## ***Scope and Limitations***

The Guidelines address only a very narrow type of research: the genre of research that evaluates the impact – that is, the effectiveness or efficacy – of a particular product or service on educational outcomes. This research is inherently a comparison of what happened with a new technology-based intervention to what would have happened if the intervention had not been introduced. Legitimate research design methods distinguish the intervention's impact from the other factors that could have influenced the results, thereby isolating it as the most plausible explanation for the impact. Decision makers may then have confidence that, if they implement the intervention in the manner described in the study, they will

receive an impact similar to that found in the research, given other limitations stated in the report.

By focusing on one kind of research in the Guidelines, we do not mean to diminish the value of other research genres to education providers and decision makers. Indeed, other research purposes, methods and designs are important to guide product design, selection and implementation. For example, formative research, in particular, has an important role in earlier stages of product 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. These questions may also be addressed by summative evaluation studies, though perhaps at a slightly higher cost. Thus providers would most often move to the kinds of evaluations addressed in these Guidelines only after undertaking a research agenda with other goals and methods.

The Guidelines will not attempt to dictate methodology or to summarize the many

volumes written on evaluation research designs. For further information on research designs, readers are encouraged to consult experts and to review such publications as Shadish, Cook and Campbell (2002) or SIIA's *Scientifically Based Research: A Guide for Education Publishers and Developers* (2003; <http://www.siiia.net>).

## **Challenges Common to Research on Educational Technology**

We begin with some perspectives and assumptions concerning common challenges related to research on educational technology products and services. While these issues are not all necessarily unique to this domain, they are challenges typically encountered in conducting such research.

### **Outcome Measures**

Because technology serves many purposes, its impact should be measured in a manner specific to the given type, goals and use of a specific technology. Student achievement and test scores are not the only valid measures. Technology purposes range from instructional to administrative, from assessment to professional development and from data warehousing systems to information productivity applications. The measures could therefore include such outcomes as student test scores, teacher retention rates, changes in classroom practice or efficiency, availability and use of data or other student/teacher/school outcomes that can be observed and measured.

Many of these outcome measures can also be viewed as intermediate outcomes – changes in practice that, as demonstrated by other research, are likely to affect other final outcomes. For example, an evaluation of a certain data system may find its positive impact on the use of data to inform instruction, and we know from other research that this outcome can help drive improvement in student learning. For purposes of the data

system, impact on the use of data is an appropriate evaluation outcome measure, and it should be valued by education decision makers as an intermediary to the ultimate goal of improved test scores.

In addition, it is essential that an evaluation study focused on student achievement select appropriate outcome measures that provide the proper balance between aligning to the specific learning outcomes addressed by the technology and providing generalizability. In some cases, an outcome measure that may be important for school accountability (e.g., state tests) may constitute too blunt an instrument to capture the full value of a certain product or service (e.g., one that is narrowly targeted such that the state test may include curriculum not covered by the intervention). Another assessment may be better aligned to, and therefore better able to measure the impact of, a product or service with a narrower or more targeted goal – including achievement on a specific set of learning standards, student technology literacy, critical thinking or student motivation. In instances when the comparison

group does not cover the material on the specialized test, then we can conclude only that it is possible to teach the material; results of such comparisons can say little about the

relative effectiveness of the product studied. In some cases, it will be desirable for the study to use both types of tests.

## Fidelity of Implementation

Research results are heavily influenced by the extent and quality of a product's or service's implementation. Educational technology implementation occurs within very complex organizational structure 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, therefore, the impact. This stands in contrast to some medical trials, where implementation variables depend less on conscious decisions of multiple actors, but are largely based upon whether subjects comply with treatment as prescribed, as well as on biological systems.

Simply providing technology without efforts to measure whether it is being used as intended and is functioning as designed – which may include the vendor ensuring necessary training, support and leadership commitment – may not

be an experimental condition that can be expected to succeed. For example, if a technology is not matched closely to the curriculum and instructional strategy, results are compromised. The condition in the research study is therefore ideally composed not only of the product or service itself, but also of the context and support for its use. In other words, the question is not simply whether the intervention works, but how well it works under particular conditions. Thus the treatment is best described in terms of an implementation model provided by the technology developer. For further information, please see the *SIIA Software Implementation Toolkit: Guidelines for Educators* (2007; <http://www.siiia.net>). At the same time, if the amount of support provided for implementation is more extensive than is normally available to customers, the evaluation may become a “hot house” study with more limited generalizability.

## Comparison Conditions

In education there is seldom a pure “blinded control” condition such as can be achieved in a medical trial with a placebo or sugar pill, where the placebo is assumed to have no effect, but the subject doesn't know whether or not it is the real medication. In schools, a new math program is typically compared to the math program already in place. Evaluations of education products and services resemble comparative effectiveness trials in medicine in which a new medication is tested against a currently approved one to determine whether it is significantly better. For any evaluation of a

product or service, the measure of effectiveness is really the comparative effectiveness against what is often called the “business-as-usual” condition. Because the effect of the product or service will depend on the existing, or baseline, way of doing things, the same product may prove effective in one district that currently has a weak program but relatively less effective in another where a strong program is in place. In both cases the technology may have a positive effect, but an impact may not register in the evaluation in cases where it is measured relative to an

otherwise effective business-as-usual condition. Thus in education, it is necessary to test products and services in a variety of settings representing differing comparison conditions. And it is not unreasonable to

evaluate a technology in a setting where there is need for improvement; potential customers seeking information on effectiveness are generally those with a problem to solve.

## **Pace of Research vs. Technology Innovation**

Technology products and services are constantly changing and improving. By contrast, in evaluation studies, several years may pass between the initial stage of identifying participants and the final stage of reporting results. In many cases, by the time research is completed, the technology products and services may have been significantly updated and no longer be available in the format or version studied. In this case, educators should appropriately consider studies conducted on previous product versions, as well as those conducted with other populations and in other settings.

Evaluation research, therefore, must be only one of many factors used for decision making. Waiting for comprehensive and definitive research literature on a given intervention will both dramatically limit educators' options and slow the pace of innovation and development. Building rigorous evaluation strategies into earlier field tests and school district pilots are approaches that can expedite product evaluation and help avoid forcing innovation and product development cycles to wait for evaluation research cycles to catch up.

## **Funding and Review of Product Evaluation Studies**

For a variety of reasons, education technology products and services face serious challenges in obtaining both funding and peer review of their evaluation research. On the funding side, relatively few outside resources are available for product evaluation studies, leaving it to the product developer to fund such studies. Government and foundation resources are limited relative to the large number of technology-based and other educational interventions calling for evaluations. And whether vendor-sponsored or funded by an independent party, research journals do not generally include studies of products beyond those that are intended to further the research literature in a given area. Education stakeholders would do well to consider these factors if they find a dearth of independently funded or reviewed product evaluation research, and to look at the merits of the available research and the product itself rather

than being unduly influenced by its treatment in the research marketplace.

In the current climate, without support from the publisher or developer, evaluation research called for by education decision-makers will not get conducted. And without non-traditional publication channels, the research that is conducted is unlikely to reach the decision-makers. An educator may find less formally reported studies that have value, and may also have the option of conducting a pilot to obtain useful evidence locally. Such pilots can often be funded through the set-aside percentage for evaluation in many federal and other grant programs that pay for the technology purchase.

Thus the formal research journal is not the only source of information about technology products and services. While maintaining a

commitment to a necessary level of rigor in conducting and reporting effectiveness research, these guidelines point to a range of

approaches that are available to vendors of technology products and services.

# Summary and Outline

The following specific guidelines comprise the core purpose of this document – to describe standards of best practice for the conduct and reporting of evaluation research on technology-based products and services. Below are the guidelines in outline format. Following the outline, the remainder of the document will fully describe these guidelines and related considerations.

## ***Ask the Right Question***

1. Match the research question (and, ultimately, its outcome measure) to the intervention’s goals. Where appropriate, include intermediary goals such as a change in practice that the research literature suggests enhances achievement or other important outcomes.
2. Select outcome measures that provide an appropriate balance between being sufficiently sensitive to the particular outcomes targeted by the product or service (e.g., a subset of learning standards), and aligning to a more general measure used for educational accountability (e.g., high-stakes state tests).
3. Before the study begins, decide whether to evaluate the product or service in one of two ways: (a) under ideal, “hothouse” conditions (i.e., an efficacy study); or (b) under ordinary field conditions, where an impact may be more difficult to discern (i.e., an effectiveness study).

## ***Support the Implementation of the Product or Service***

4. Develop and document as explicit a model as possible for how to implement the product or service in the educational setting. This includes the appropriate technology infrastructure, educator training and product usage required for an impact to be detected. The more explicit this model, the more likely the research will be able to explain the results in terms of whether the implementation met these expectations.
5. In conducting an evaluation, distinguish between correlational and causal findings. It can be useful to check for correlations in the data, such as between implementation fidelity and outcomes, for the purposes of product improvement and understanding best practices. But be cautious in drawing conclusions about a causal effect of the intervention from correlational findings, as a factor other than the intervention can often provide a plausible explanation.

## ***Plan a Study of Sufficient Size and Duration to Demonstrate an Effect***

6. Establish the study’s “unit of analysis.” This is the sample unit level – typically school, teacher or student – at which the product or service is designed to be used. The appropriate unit may be determined by the implementation model, as when the model requires treatment to be administered school-wide. Otherwise, the unit of analysis may be determined based on cost constraints. For example,

it costs less to randomize 40 students than 40 schools to treatment and control conditions.

7. Employ a sample size sufficiently large to draw conclusions with statistical confidence, taking into account the magnitude of the expected effect, the availability of a pretest and the number of units of analysis needed.
8. Plan for a study of sufficient duration for the product or service to have its intended effect. Consider the period needed for training and other start-up activities, and allow time for full integration into instructional and administrative processes.
9. Identify the comparison condition or clearly defined baseline relative to which the estimated impact of the evaluated product or service is measured. The comparison condition is needed to determine what would have happened without the new product or service.

### ***Plan for Plausible Causal Claims***

10. Choose a research design that is capable of reducing plausible alternative explanations for changes in performance, other than the impact of the product or service under study.
11. Avoid or mitigate selection bias in identifying the group that uses the new product or service and the group to which it is compared. A method to be considered is random assignment of study units (e.g., school, teacher or student) to use the intervention. Where random assignment is not feasible, other approaches to identifying a well matched comparison group can be used.

### ***Avoid (the Appearance of) Conflicts of Interest***

12. Follow standards of practice and regulations put in place to protect the privacy and safety interests of study participants. These often include review by an Institutional Review Board and adherence to the Family Educational Rights and Privacy Act (FERPA).
13. Work with researchers who can be objective and independent. Take steps in selecting the researcher, determining the editorial and reporting process and funding the study that will help ensure objective findings. This applies whether the research is conducted internally or through an external contractor.
14. Design participant incentives to avoid any bias in the results. 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.

### ***Provide a Comprehensive and Detailed Research Report***

15. Produce a full research report that thoroughly describes the research conditions and context in detail, including the product or service, its implementation, group

assignments, comparison conditions, populations, interactions and any factors that may cause bias. Only a sufficiently detailed report allows for a third party to evaluate its conclusions and, potentially, to replicate the study.

16. Distinguish between (a) the findings pertaining to the original core hypothesis and (b) the exploratory results and conjectures arising from post-collection review of the data.
17. Be clear about the study origins, initiating parties and funding sources.
18. Be clear about study authorship and final editorial control.

### ***Make the Research Findings Widely Available***

19. Make the research report available through a variety of channels, such as a refereed (peer reviewed) journal, conference presentations, research clearinghouses and the company website.
20. Make all formal evaluation research findings available upon request regardless of the outcome, except in these instances: (a) a “failed experiment” where it is determined prior to review of outcomes data, for example, that the product or service was not implemented with fidelity, too few participants could be recruited, the study was too poorly designed or the data could not be collected; or (b) determination by the provider that the product or service must be improved and re-released, in which case the results can be considered as formative information for product improvement.

### ***Accurately Translate Research for Customers***

21. In the marketing literature for a product or service, accurately describe its impact – relative to the strength of the research design, quality of the evidence and size of the effect – using language that educators without research training can understand.
22. Cite the full research report any time the research or its findings is referenced.

# Guidelines and Considerations for Conducting and Reporting Product Evaluation Research

The following sections elaborate on the Guidelines outlined above and discuss related considerations. This full detailing comprises the core purpose of this document. Although, as noted in the introduction, the Guidelines do not provide a textbook on research methods, each is supplemented here with additional recommendations for how to address challenging issues specific to its implementation around research that evaluates the impact of educational technologies.

## ***Ask the Right Question***

For purposes of these Guidelines, evaluations of educational technology products and services are designed around a narrowly defined question – whether a product or service has an impact compared to what would have happened if the product or service were *not* put into service. There are many ways to approach this question, the choices of which will be critical both to the study’s chances of success and to the value of the research and its findings for various audiences. It is important to establish the specific questions, how they will be answered and the criteria for a successful answer before data collection begins. Fishing after the fact for interesting results that might simply be a matter of chance is avoided by an explicit research design and protocol document.

- 1. Match the research question (and, ultimately, its outcome measure) to the intervention’s goals. Where appropriate, include intermediary goals such as a change in practice that the research literature suggests enhances achievement or other important outcomes.**

A partial inventory of educational technologies consists both of management and of instructional technologies, the latter including those designed for students, educators or both. A partial list contains student information systems, curriculum management systems and professional development programs as well as instructional software, digital content, learning tools and assessment applications. Because of the variety of applications, and because a given evaluation can focus on only a limited number of outcomes, it is essential from the beginning to specify exactly what questions are most worth answering. Simply put, “Does the technology work?” is a sufficient question only when we add details addressing the larger question, “and to what end does it work?”

Nothing inherent in the methodology of evaluation research restricts it to a focus on student performance on high-stakes tests. This is particularly relevant for education technologies, which may be able to accomplish multiple goals. Thus a study can be concerned with effects of an intervention on skills that either differ from, or that overlap with, what a given standardized test measures.

Moreover, even when improved student achievement, however defined, is the ultimate goal of a product or service, an evaluation of their impact could instead (or in addition) measure desirable intermediary outcomes or changes in practice. For example, student information systems may increase the availability of data needed for classroom and

school decision making, professional development programs may improve teacher

skills and retention and instructional software may drive small group instruction.

**2. Select outcome measures that provide an appropriate balance between being sufficiently sensitive to the particular outcomes targeted by the product or service (e.g., a subset of learning standards), and aligning to a more general measure used for educational accountability (e.g., high-stakes state tests).**

Once we understand where the product or service is intended to have an impact in terms of research questions, the next step is to select appropriate outcome measures. As noted, the technologies tested by research are typically aimed at improving student achievement. In such cases, a tension may exist between detecting what may be an intervention's small effect on a very general measure (e.g., a state test), and showing that it has a larger effect on a measure that is more sensitive to the specific set of learning standards/goals the intervention is designed to address. State tests and nationally standardized tests are often very general and may have relatively few of the items associated with the specific goals of, for example, the targeted supplemental product or service being evaluated. There is a similar tension related to the variation among state learning standards and assessments.

Where student achievement is the goal, it is most helpful to use an outcome measure – including one composed of multiple assessments – that:

- has widespread credibility and acceptance;
- includes sub-strands that are closely aligned with the particular product or service goals; and
- measures growth over time along a consistent scale.

Using such a measure will help ensure that we estimate the impact of an intervention on a range of skills that underlie a general construct, typically summarized by the results

of standardized tests, and will also help ensure that more fine grained results pick up impacts congruent with the particular focus of the product.

Where the intended impact of the product or service is targeted to a very specific skill, it may be necessary to create a custom test or other measure (e.g., a protocol for classroom observations). 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 in some cases be necessary, doing so carries a risk of unreliable or invalid results. Even when well designed, such tests may do a good job of measuring the specific outcome but may not tell the educator whether the product or service is likely to make a difference on a more broadly based measure (such as a high-stakes test). Moreover, in instances where the comparison group has no exposure to the concepts underlying an instructional product or service being studied, a test specifically of those concepts will have little meaning. The provider should work with the researchers to identify outcome measures that will be meaningful metrics for the research audience, given the kind of product or service being studied. More than one measure may be needed in order to capture both specific skills and broader constructs.

If a test is used that is not standardized, the researcher should provide information concerning the test's reliability and validity or give a reference to the technical documents describing characteristics of the test. If a custom test is developed, the researcher should examine and report characteristics of the test,

including an index of internal consistency and how well the test correlates with other established measures. If performance assessments are used that involve judges' ratings of performance, the consistency of those ratings should be reported.

**3. Before the study begins, decide whether to evaluate the product or service in one of two ways: (a) under ideal, "hothouse" conditions (i.e., an efficacy study); or (b) under ordinary field conditions, where an impact may be more difficult to discern (i.e., an effectiveness study).**

Evaluation research is often divided into two general types. There is a distinction between *efficacy* or "hothouse" studies that demonstrate how a product or service works under ideal conditions and *effectiveness* studies that test it on a larger scale under regular field conditions.

In the efficacy study, the researcher would monitor implementation and intervene to ensure delivery of the training, infrastructure, support and so forth required to nurture the intervention and to ensure that it is implemented as recommended. Thus efficacy studies show how a product or service works when used as intended. By contrast, an effectiveness research model would provide no more than the usual level of training and support that an ordinary customer would receive. Both types of studies are legitimate approaches to evaluation.

Efficacy studies are often conducted on a smaller scale, with fewer schools and teachers, in part because of the cost of monitoring and of providing extra support to each participant. Because of the extra support, the impact may be much larger – meaning, from a research design point of view, that fewer teachers or schools are needed to detect a difference in outcome using a statistical test. A hothouse 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.

While large effect sizes can be more difficult to obtain in an effectiveness study, the results are more meaningful to others, as they reflect the ordinary conditions of implementation.

### ***Support the Implementation of the Product or Service***

The implementation of educational technology occurs within a very complex organizational system. In the school, classroom or virtual learning environment, this system requires not only the products and services but also provision of resources such as computer systems, training, planning and classroom time, as well as sound use in terms of the technology's integration into the curriculum and pedagogy. This set of guidelines assumes that the intervention, as applied, consists of more than the packaged product or service, and that, therefore, the research must consider how the product or service is used.

**4. Develop and document as explicit a model as possible for how to implement the product or service in the educational setting. This includes the appropriate technology infrastructure, educator training and product usage required for an impact to be detected. The more explicit this model, the more likely the research will be able to explain the results in terms of whether the implementation met these expectations.**

Because implementation is related to impact, an explicit model for how to implement the product or service provides a benchmark against which the quality and quantity of the implementation can be measured. Presumably, the model would have resulted from prior research and pilot testing of the intervention in educational settings.

The model should document, for both customers and researchers, the following elements, as appropriate:

- the professional development required;
- the amount and type of technology infrastructure needed;
- the level and type of support the technology will call for;
- 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.

For related information, see “*SIIA Software Implementation Toolkit: Guidelines for K-12 Educators*,” April 2007.

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

With an explicit model, it is then possible to specify “fidelity of implementation” or the set

of conditions under which the provider predicts that the product will have the greatest effect and, therefore, the conditions that the customers or research participants are expected to put in place. Although not guaranteeing success, an explicit model makes it possible for the researcher to monitor and document the degree of implementation fidelity. While it is incumbent on the company 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.

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, the ability of an experiment to detect a difference increases when implementation is consistent across all the schools or classrooms using the product or service. Ideally, the provider works with the education customer to ensure that all the necessary support and training are provided and that other conditions are met. In general, in the case of effectiveness evaluations, it is not the role of the researchers to support the product or, in most cases, to provide feedback to the provider when implementation is failing. Given an explicit model, the researcher documents both the support provided and extent to which the implementation model was achieved.

- 5. In conducting an evaluation, distinguish between correlational and causal findings. It can be useful to check for correlations in the data, such as between implementation fidelity and outcomes, for the purposes of product improvement and understanding best practices. But be cautious in drawing conclusions about a causal effect of the intervention from correlational findings, as a factor other than the intervention can often provide a plausible explanation.**

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. It is tempting to look only at the effects found for the teachers or schools that implemented the product or service fully. It is also tempting to examine the correlation between the extent of implementation and the outcome being measured. However, researchers recognize both these strategies as having the potential for misleading conclusions. (This issue will be addressed in terms of research design in guideline 10.)

This potential is easily understandable. Improved performance that is correlated with stronger implementation may be due to characteristics of the strong implementers and not the intervention itself. Teachers who are better implementers may have found the product or service to be more interesting or may simply have been more energetic and willing to try something new. The results they achieve may therefore reflect their general

enthusiasm instead of the effects of the intervention. (This example illustrates a fundamental challenge for research design that is addressed in later guidelines: how to separate the impact of introducing a new intervention from other characteristics of users, schools or student populations where it was tried.)

At the same time, correlational data can be very useful in other ways. By looking at the most successful users, 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. In this way, evaluation research can also be used for formative purposes, and this two-for-one design can be an efficient use of research resources. After-the-fact exploration of the results for these kinds of relationships can also help shape the next round of research, when more refined hypotheses can be tested.

### ***Plan a Study of Sufficient Size and Duration to Demonstrate an Effect***

This set of guidelines touches on some basic considerations in research design, but does not go into technical detail. They are based on the assumption that evaluation research makes use of methodology and statistical tests to determine whether the measured difference between two groups in a sample is probably a real difference caused by the intervention or just a chance occurrence. The larger the intervention's impact and the larger the sample in the experiment, the more likely it is that an effect can be confidently detected through the uncontrollable "noise" of the research setting. As cost considerations push toward smaller, shorter term experiments, it is important to strike the right balance.

**6. Establish the study’s “unit of analysis.” This is the sample unit level – typically school, teacher or student – at which the product or service is designed to be used. The appropriate unit may be determined by the implementation model, as when the model requires treatment to be administered school-wide. Otherwise, the unit of analysis may be determined based on cost constraints. For example, it costs less to randomize 40 students than 40 schools to treatment and control conditions.**

Determining the unit of analysis – essential for deciding on the size 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 be independently treated?” For a home-based tutoring system sold on a consumer basis, the smallest unit may be an individual student. But most educational technologies are designed to be used in group settings with all students in a classroom. Formative assessment systems may be designed for a whole school and call for leadership training and school-wide support for implementation. Course management systems may be implemented district-wide.

The level of implementation will help determine the size and cost of a study. The number of units needed for an experiment, as determined by a power analysis, is similar whether the unit is constituted as an individual student, a class, a team of teachers at a grade level, a school or a district. For example, if the technology is implemented and its impact measured across a whole school, as in a school-wide reform, we may need 40 schools and many thousands of students for the study. If the intervention is provided to each student independently, as in instructional software accessed via a login with students assigned at random to use the product, we may need fewer than 100 students. The larger units will naturally involve far more individuals – and cost – than the latter.

In some cases, the implementation model can adapt to the need 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. 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.

There are two related arguments against artificially dividing up the normal unit of implementation. First, there is a danger of what researchers call “contamination.” For example, where 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 or service, 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 otherwise would. For example, dividing teachers within a grade level may interfere with informal professional collaborations that otherwise might support teacher effectiveness. Similarly, if only a few teachers in a district use the intervention as

part of a larger multi-district experiment, this presence may be insufficient to gain the administrative/technical support, training or leadership needed for implementation, or the commitment of the participating educators to make changes to their curriculum and instruction necessary for success. 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.

## **7. Employ a sample size sufficiently large to draw conclusions with statistical confidence, taking into account the magnitude of the expected effect, the availability of a pretest and the number of units of analysis needed.**

To a large extent, the relative cost of a study within a given research design is related to its size, especially when extensive data collection is involved. “Under powering” the study can lead to inconclusive results. The previous guideline illustrates some of the complexity in identifying the unit to be counted when determining a sufficient size for a study. A statistical process called a “power analysis” allows researchers to predict approximately how many units (whether students, teachers, schools or some other unit) typically will be needed for reaching a conclusion in which they may have confidence.

Several factors are important to consider when determining sample size.

- The magnitude of the effect, or the intervention’s effect size. How large an effect is the product or service expected to provide? Specifically, how big a difference is expected at the end of the study between the group that received the intervention and the comparison group? If a large impact is expected, a smaller study can detect it. Conversely, if a smaller impact is expected, a larger study may be needed.
- The use of a pretest. A measure of pre-intervention performance is perhaps the most useful study element to help limit

its size. The benefits of the pretest come from its utility for predicting posttest performance independently of an intervention’s effects. The pretest need not be the same test as the outcome measure or even in the same subject; it simply has to be correlated with the outcome. For example, a reading score can be used as a pretest for a science outcome. However, the stronger the correlation, the greater the benefit. Including other variables associated with the students, teachers and settings in measures of pre-intervention performance can also have substantial benefit.

- The size of the unit of analysis. As noted above, an experiment that uses schools as its basic unit of analysis will, de facto, be larger in terms of numbers of teachers and students than an experiment that uses teachers as its basic unit. Guideline #6 addressed the tradeoffs in choosing an efficient yet appropriate unit of analysis.

Finally, it would be wrong to conclude that an intervention had no effect just because the impact did not reach statistical significance. A finding of no significant difference or no discernable effect may occur because of an insufficient sample size. A small effect may

be educationally significant, even if not statistically so.

**8. Plan for a study of sufficient duration for the product or service to have its intended effect. Consider the period needed for training and other start-up activities, and allow time for full integration into instructional and administrative processes.**

Study duration is an important consideration for any study or design. Generally, the longer the period of time the product or service is in use, the greater likelihood that its effects will be measurable. There are two areas of concern with duration: (1) obtaining 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 involved takes time for participants to absorb. In such cases, 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 for a product or service, once fully

implemented, to have an impact. When students take considerable time to move through the educational content, the full year of implementation is likely needed. Conversely, if the content to be learned is quite specific and limited in scale, then the length of the study can be much shorter. A concern with a very short study is that 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 shorter 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.

It is also possible to consider a study as consisting of 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.

**9. Identify the comparison condition or clearly defined baseline relative to which the estimated impact of the evaluated product or service is measured. The comparison condition is needed to determine what would have happened without the new product or service.**

The basic logic of effectiveness studies is a comparison of two or more conditions. In education, there is seldom a placebo or condition designed to have no impact, and so

the comparison is usually between a treatment group using the new intervention and the comparison group using the business-as-usual program that is already in place. In this case,

the comparison condition serves as a measure of what would have happened to the treatment group if they had not been provided the new product or service. Note that some studies involve head-to-head comparisons of two distinct interventions that are both being tested, rather than comparison to a business-as-usual condition. Thus the size of the impact of the intervention being tested is always relative to the effectiveness of the other, perhaps pre-existing, program. The question, then, isn't simply how much growth occurred in the treatment group; instead we ask how much more growth occurred there in comparison to what happened in the other group. The impacts that researchers are looking for are generally quite small.

The idea of “relative effectiveness” also suggests that the comparison of a new intervention to one that is quite effective will yield a result different from a comparison to one that is less effective. Fortunately, the

process of recruiting interested research sites can often provide a fair and realistic approach. Specifically, a potential research site that has a well functioning solution to the problem that the new product or service addresses is less likely to volunteer. Conversely, a potential research site will more likely volunteer if decision makers are interested in solving a problem they have otherwise failed to address. Thus, all other things being equal, the research sites that volunteer often provide the best candidate for a relevant contrast between the treatment and the comparison conditions. Note that this strategy will sometimes create conflict between research and sales goals, since the most appropriate research site may also be a good sales prospect. Joint planning can often help support sales without inappropriately interfering with the research design, and a pilot implementation may sometimes serve the interests of both research and sales.

## ***Plan for Plausible Causal Claims***

Effectiveness research seeks to isolate and measure the impact of a product or service and to prove that this intervention, and not something else happening concurrently, caused the observed change in performance. A demonstration of this sort would be easy if researchers had time machines. After observing what happened to a group without the intervention, they then would go back in time, provide it and measure the difference in performance. Everything remains the same except that, the second time, the new product or service would be in use. Lacking time machines, we have to provide the intervention to a group and find some way to determine what would have happened if we had not done so. This is the core of evaluation research design. The following guidelines present these ideas in common sense rather than technical terms.

### **10. Choose a research design that is capable of reducing plausible alternative explanations for changes in performance, other than the impact of the product or service under study.**

There is a hierarchy of designs and techniques – from weak to strong – that can be used to study effectiveness and demonstrate causal impact (i.e., the intervention caused the outcome, as opposed to the outcome being caused by another factor that happens to be

correlated with the intervention). The goal of these methods is to remove competing causal explanations (called confounding variables) so that the intervention under study can be credited with any observed changes.

It is important to point out that there are many ways to do this, including:

- Looking at the group's performance before and after introduction of the product or service without any comparison. This very rudimentary approach could have value where the assumption can be made that, without the product or service, there would be no growth. This might be an appropriate research design for a product or service involving very esoteric content
- Measuring successive cohorts. This approach provides a measure of past group performance. An example would be determining a baseline in the average math scores for prior cohorts of eighth graders and then measuring the math tests of cohorts after the introduction of a new math program. This design requires knowing the seventh-grade math scores for each student in the study.
- Comparing the performance of the group receiving the intervention to the performance of a group that is similar in many other respects but not receiving it.
- Randomly assigning some in a group to use the intervention and others to continue with business as usual, as is done in a randomized control trial, generally considered the strongest method for causal inference.

An otherwise relatively weak design such as simply measuring a difference between, for example, last year's seventh graders and this year's seventh graders, can be useful if other explanations for changes, such as environmental or social conditions at the research site, can be eliminated. Stronger methods, relying less on this local knowledge and more on research design, ensure that plausible rival hypotheses can be ruled out.

The following examples are used to illustrate weak to strong research designs:

In a very simple successive cohort design, the principal of a school may observe a good result – say, the school's percentage of proficient students in a particular grade increases over the previous year – after having implemented an educational technology, and the researcher would ask whether other things also happened that could explain the improvement. Were there other changes to the curriculum or instruction? Was the test rescaled? Did new teachers join the school? Did the boundaries of the school neighborhood change? These and many other questions are quite reasonable and, as a practical matter, 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, although there may still be other plausible explanations the principal has not considered.

In a second case, a district decision maker may be considering the results for 20 schools that implemented the intervention. The results are generally good, but in some schools there has been turnover of the principal and, in many schools, large numbers of new teachers make the pattern less clear. At this point, the researcher may suggest a comparison involving another 20 district schools with similar characteristics that did not use the intervention. A study might investigate whether, on average, the 20 schools that chose the new technology performed better than the ones that did not. A statistical test is used to determine whether the results in the intervention schools were significantly different from the others. But a statistical test does not eliminate other plausible explanations for the difference. Perhaps the leadership and morale in the intervention schools were better,

resulting in both an interest in the technology and better results. To some extent, statistical controls can be applied to minimize confounding variables' effects. For example, using a measure of teacher experience or even of teacher morale, researchers can mitigate the effect, essentially removing it as an explanation for the differences between two groups of schools. But there are limits to statistical equalization. When the two groups start out very dissimilar – for example, if the district's lowest scoring schools are the ones receiving the new product or service – it isn't possible to adjust for these baseline differences that constitute a plausible explanation for performance differences.

In a third case, unlike other research designs, randomized control using a sufficiently large sample eliminates most other plausible explanations by providing teachers or schools with the product by chance rather than through personal choice or confounding characteristics. Interest and morale play no role in assignment of the schools to use the new technology. The researcher's estimate of the impact of the technology is not biased by these or any other characteristics, although the two groups may still be unbalanced just by chance. It is useful also to note that a randomized experiment is generally simpler to design, requires less information and is easier for the researcher to analyze, all of which should translate into a lower cost.

**11. Avoid or mitigate selection bias in identifying the group that uses the new product or service and the group to which it is compared. A method to be considered is random assignment of study units (e.g., school, teacher or student) to use the intervention. Where random assignment is not feasible, other approaches to identifying a well matched comparison group can be used.**

This guideline is an extension of the previous one, elaborating on approaches to finding well matched groups to compare. 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. Because of their interest and, perhaps, their enthusiasm as volunteers or early implementors, such schools or teachers may exhibit other strengths that can reasonably provide an alternative explanation for differing results. In other cases, when schools or students most in need are provided with the intervention, it is difficult to compare them to other schools within their district. Even if one attempts to compare the intervention schools with similar ones in other districts, a potential bias is introduced, in that the many possible discrepancies between the districts could

account for any differences in performance found between their schools. (Note: A methodology called “regression-discontinuity” can take advantage of this type of assignment of schools into conditions; however, the assignment should follow strict criteria – a condition that is seldom obtained in school systems.)

Although many methods for mitigating these biases exist, there is one recognized sure-fire way to do it: random assignment. In educational research, the units that are randomly assigned are seldom individual students. More often, the units of assignment are teachers, schools or grade-level teams within schools. Starting with a large group of teachers (or other units) and randomly assigning half to a treatment or intervention group and the others to a control group means that the only systematic difference between the

two groups is that, for some, the coin toss came out heads; for others, tails. Because the two groups are basically the same, the control group can represent the intervention group's likely performance level, had it not received the intervention.

There are impediments to random assignment in education. A practical consideration is that it requires concurrently planning 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. Ethical concerns are also sometimes raised when group assignment is viewed as depriving some equally needy students of the new product or service. However, several arguments support its use. For example, in cases where there is limited availability of the intervention or where it is being rolled out in phases, a lottery may actually be the fairest method of

distribution. Second, some methods ensure equality by providing the control group the intervention either after the study is completed or in a later stage of the study. It can also be argued that, because it is unknown whether the impact will be positive, the control group is not necessarily being deprived of a benefit.

Where random assignment is not feasible, a wide variety of quasi-experimental methods such as those described in the previous guideline can construct comparison groups to statistically match a treatment group, either before or after the fact. With any of these techniques, there remains the possibility that the result may be biased by some characteristic that wasn't measured and controlled for. Still, because it is often more feasible to carry out, because large amounts of evidence may be accumulated and because the result may be of a quality that educators find useful, a well controlled quasi-experiment may also be a good approach.

### ***Avoid (the Appearance of) Conflicts of Interest***

The next set of guidelines addresses a very different kind of potential bias that provider-sponsored research on a product or service should be especially sensitive to addressing. Doing so will help to ensure that the conduct and reporting of such research is objective and provides results that can be trusted.

#### **12. Follow standards of practice and regulations put in place to protect the privacy and safety interests of study participants. These often include review by an Institutional Review Board and adherence to the Family Educational Rights and Privacy Act (FERPA).**

For much of the evaluation research conducted in schools, the classroom experience falls within the normal expected educational activities, and there is no appreciable risk to students or teachers. Nevertheless, strict confidentiality and "human subject" protections apply. Two major sources of legal and ethical standards help researchers avoid potential liabilities and demonstrate an

intention to conduct research for the public good.

The Family Educational Rights and Privacy Act (FERPA) sets out 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 a practical and 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 for research where the purpose is to improve education and the identity of the student remains confidential. If the goal of the research is simply to generate marketing statements and 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: This regulation is complicated, and neither this summary nor anything else in these guidelines should be viewed as legal guidance.]

Evaluation researchers should also take steps to observe the ethical standards established within the research community for conducting research using human subjects. Research organizations, whether universities or for-profit firms, must obtain approval of an Institutional Review Board (IRB), which

reviews proposed research procedures and determines whether the activity constitutes legitimate research and whether there are appreciable dangers to participants. In the case of IRB review, the issue is consent to participation in the research, not consent to the release of records.

An important principle in ethical 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 about an implementation that is already underway, the selection of intervention 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 in the hopes of persuading the teacher to stay in or to withhold other promised payments.

**13. Work with researchers who can be objective and independent. Take steps in selecting the researcher, determining the editorial and reporting process and funding the study that will help ensure objective findings. This applies whether the research is conducted internally or through an external contractor.**

Provider-sponsored evaluation research can be conducted internally or by an external contractor. In either case, explicit steps should be taken to prevent undue influence (and the perception of it) and to help ensure objective and independent findings. These steps are taken before work begins. Following are the primary examples of such steps:

- Create a clear separation of the provider's internal research function aimed at producing publicly available evidence of effectiveness and the marketing/communication functions. [Note that this step does not include formative product testing as part of a continuous improvement model, which need not be separate from marketing or product development.]

- If the provider’s internal research staff conducts the study, credibility is enhanced if they can report results regardless of outcome and legitimately publish their own reports. Credibility is greatly enhanced if the report can state that the researcher was given autonomy to publish before the study’s data were collected.
- If an independent research organization conducts the study, the credibility of their independence is enhanced if they are given authorship and editorial control with a distribution license to the provider. Giving them final editorial control also enhances the researcher’s independence by providing additional motivation to later seek publication.
- If possible, secure third-party involvement (which could include funding) such as from a government agency, foundation or educational institution.
- Submit the study’s report for independent review.

research design and analysis and 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 primarily 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.

An independently funded outside research group provides the strongest assurance of objectivity. In this model, the provider assists an outside researcher in obtaining a grant from a government or foundation. Such funding often requires an existing body of independent studies and tends to be highly competitive. Unfortunately, this is seldom a viable option, as the supply of funding is limited relative to the demand in terms of the many products and product studies.

Many education technology providers employ qualified researchers who have expertise in

**14. Design participant incentives to avoid any bias in the results. 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.**

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 as an inducement to participate. Also common is the practice of providing a modest honorarium for efforts required beyond regular classroom work. Districts are often interested in participating because a study offers 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.

Even if there is no perception that the institution/educator benefit depends upon a study’s results, inducements do have the potential for bias in some situations. It is therefore 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. Also, as previously stated, it is appropriate to offer the intervention (and training) to the comparison group once the study is over.

When a teacher is not implementing an intervention under study with fidelity, providing extra support (e.g., professional development) beyond that normally provided to any educator outside a research setting is appropriate only in efficacy studies (and not in effectiveness studies), where the extra support

would be reported as such. As described in other guidelines, for effectiveness studies, it is important to restrict support and training to that which would normally be provided to customers not involved in research studies.

It is worth noting that provision of free materials and generous incentives can sometimes be de-motivating. In the best situations, the school is invested, literally, in the product or service and therefore in the research, and so is 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 or with complying with the research requirements to carry out the research properly.

### ***Provide a Comprehensive and Detailed Research Report***

While there will likely be several versions of a study's report, depending on the audience, all should build from and link back to the comprehensive research report.

#### **15. Produce a full research report that thoroughly describes the research conditions and context in detail, including the product or service, its implementation, group assignments, comparison conditions, populations, interactions and any factors that may cause bias. Only a sufficiently detailed report allows for a third party to evaluate its conclusions and, potentially, to replicate the study.**

It is important that the original study be reported in sufficient detail so that (1) other researchers can potentially replicate the original study and ultimately confirm or disprove its findings and (2) educators can compare the study conditions with their own local conditions and estimate what would happen if they were to implement the product or service themselves, or what changes and associated costs would be needed to replicate the implementation. In addition, researchers intending to combine studies on the same product or service into a research synthesis

need the full technical information typically contained in a full research report.

The paragraphs that follow contain important elements of a research report that are often overlooked. Without these elements the report is less useful and should be viewed with caution.

- The report should clearly describe the intervention. This description includes not only the core product or service itself, but also the training and support delivered by

the provider, the technology hardware and infrastructure, additional support and resources from the school or district, the overall instructional program and practices as amended by the intervention and other relevant features. The amount of support in comparison to a typical implementation should be described to enable readers to determine their ability to replicate the implementation. Where setting and implementation vary widely across the study participants, it is important to indicate the range and variance of support and other factors, relative to the suggested implementation model (see guideline 4).

- The report should identify the version of the product or service that was implemented. This is especially important where there is a significant lag in publishing the report, and the intervention has been improved in important ways (or replaced entirely with a new version). While the report may not apply directly to the new version, a detailed description of the intervention that was used, highlighting its essential features and strengths, will allow educators to determine the report's current relevance to their new version. If a study on an earlier version is cited to provide evidence of effectiveness for a later version or product, then the differences between the two should be reported, perhaps by adding a postscript or addendum to the original study.
- The report should detail what the comparison group was doing, including the curriculum, program and practices that were used. While another educator testing the product or service would not attempt to replicate the comparison condition, it is important to know from what baseline the impact of the intervention was calculated. When a study is conducted across a large number of school systems, each with its

own conditions and existing programs, it may be impossible to fully document the comparison condition. In that case, it is important to identify the products or services to which the intervention was compared.

- The report should provide key data for both the intervention and comparison groups, including the student demographics and average pretest scores and the range of teacher experience, among other data. A set of such sufficiently detailed factors can enable decision makers to recognize how similar the research setting and population is to their own.
- The report should address not only the main outcomes of the study, but also the secondary or exploratory results. The main outcomes may consist of estimates of the average difference between the intervention and comparison groups. Secondary or exploratory results will often include estimates of interaction effects, which measure the extent to which an intervention is differentially effective based on other variables unique to some participants. For example, an overall result of no discernible difference between intervention and comparison groups often masks a finding that the intervention worked very well for some part of the population but not for another, or that more experienced teachers may take better advantage of it. This can be useful information, both for educators and for providers, in understanding how best to implement the intervention.
- The report should disclose factors that may indicate undue influence. As described below and in other guidelines, these disclosures include an explanation of site selection, recruitment of and incentives for

participants, loss of participants (attrition) and choices of outcome measures.

- The report should detail and explain the loss of study participants. Attrition, especially if it occurs more in one condition than the other, may be a sign that the study has been biased. For example, if a product or service results in more low-scoring students dropping out in the intervention group than in the comparison group, the outcome will be affected by the attrition. In other cases, where attrition can be documented as unrelated to the

experiment, no harm is done to the study's conclusions.

- The report must be clear about the limitations of the study, especially with respect to generalizability. All studies have limitations, and presenting these limitations clearly is necessary if readers of the report are to understand how to use the research for decisions within their local context. The limits to generalization must be stated in relationship to a full description of the sample, the comparison condition and the characteristics of the students, teachers and setting.

## **16. Distinguish between (a) the findings pertaining to the original core hypothesis and (b) the exploratory results and conjectures arising from post-collection review of the data.**

As was stated in guideline 5, researchers must specify in advance of the study where the impact is expected to be found and why, and they must include this hypothesis as the basis of the study report. Ideally, researchers in planning the study identify one or two factors that they believe are most likely to be impacted by (or that will moderate the impact of) the intervention and limit their firm conclusions to those outcomes. As the number of observed outcomes increases, the chances also 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 primary. 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.

This does not mean that researchers must limit the number of outcomes that they decide to examine before running the analysis, nor should they be prevented from identifying additional outcomes to explore after performing the planned analyses. In fact, much can be learned by inspecting the patterns of results and identifying surprising relationships. Such examinations include checking to find whether a correlation exists between quality of implementation and outcome. However, most importantly, analyses of this kind are considered exploratory, and firm conclusions should not be drawn from them. When conjecture or exploratory findings are included, they should be labeled as such.

Still, exploratory research findings are not mere statements of opinion. Research reports should provide as much rich detail and conjecture as will be useful for educators implementing the intervention. These conjectures are also important for other researchers or for planning the next set of

studies, where they may become hypotheses to be tested.

## **17. Be clear about the study origins, initiating parties and funding sources.**

There should be a clear statement about the reason the study was initiated. Because bias can enter into a study through multiple paths, a disclosure and explanation of the initiators and their purposes is important. For example, did the original request for the research come from a customer or a potential customer and, if so, what were the circumstances or decision contexts for initiating the study? Commonly, the provider initiates the study for the purpose of acquiring generalizable evidence of effectiveness that can be presented to potential customers. In this case, several school systems may be recruited as participants. In other cases, a school district may have purchased the intervention and then later invited the provider or another entity to conduct research.

The funding source for the research study must also be reported. Often this will be the provider, but it may also be a foundation, government agency, a school district or some combination. Similarly, the funding source (including in-kind contributions) for the intervention and related infrastructure, training and support must be reported. Again, sources

may be in combination, with a research grant covering the intervention, the provider donating training and the school providing the infrastructure and support. Crediting the in-kind contribution of the school systems, which often invest significantly in central office and IT staff time, is appropriate. It is worth noting that at least partial investment by the school system helps ensure the support of the district administration for a robust implementation.

As noted in guideline 14, the incentives, if any, provided to the participants may constitute a potential bias. For schools and teachers, these may include receiving the intervention at no cost, funding for training time and honoraria or other payments. Given that differential enthusiasm resulting from generous pay to the intervention group can bias even a randomized experiment, disclosure of the extent of these incentives is relevant. It is also relevant to the outcomes whether adequate release time or other supports were provided. While dollar amounts do not need to be reported for intangible resources, it is relevant to indicate how participants were rewarded and who provided the rewards.

## **18. Be clear about study authorship and final editorial control.**

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. Previous guidelines address bias and conflict of interest – or their perception – relative to the sponsorship and conduct of research. This guideline applies these issues to reporting.

Where the contract with the researcher takes the form of a “work made for hire” or where the provider retains the report's copyright, readers may assume that the provider has final editorial control unless specifically stated otherwise. Several alternatives may help to eliminate or mitigate any such perceptions:

- The study report can be attributed to the external author (recognizing that

this action alone may not be perceived as evidence of independence of final editorial control).

Assigning the copyright to external author or research organization is the clearest way to indicate that the provider is not exercising editorial control.

- If the report will be authored by several individuals with different affiliations – for example, a research company, a school system and the provider itself – it is generally assumed that the first author has final editorial control. Still, this should be clearly stated.
- Where the research was conducted collaboratively or through a contractor employing subcontractors (e.g., for data collection or observations), several authors may be credited. In such cases, the roles of the different entities should be explained, perhaps identified by the report’s acknowledgements or foreword.
- The entity with final editorial control should always be clearly identified. Reviewers of research will need a

specific contact who can speak for the methodology and results. This person will need the authority to revise the report as well as to assign copyright as necessary to a scientific journal.

- When the provider allows internal researchers to report results regardless of outcome and to publish their own reports, this policy should be clearly stated in the study.

Finally, it should be noted that the individual or entity with final editorial control is not relieved of the duty to review the research results and report with the provider prior to publication. Misunderstandings of the product or service’s characteristics or goals may need to be clarified, and all sponsor questions should be answered before making the research public. Similar policies are strongly recommended for any party conducting either an initial evaluation of a provider’s intervention or conducting a later review of such research.

## ***Make the Research Findings Widely Available***

An expectation in the scientific community is that research findings are made available regardless of the result. This does not always happen, because researchers and research journals tend to prefer reporting those studies with significant positive results. Now, however, with the establishment of research review services and web-based clearinghouses, and with the ability to self-publish or post online, this “gray literature” – never before formally published – can be more easily made available.

### **19. Make the research report available through a variety of channels, such as a refereed (peer reviewed) journal, conference presentations, research clearinghouses and the company website.**

Peer review is common in the scientific 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.

One such method is having a research report published in a professional journal. It often provides a high level of credibility, because the work has been reviewed carefully by other researchers. However, such review tends to be a lengthy process, and requires a commitment of effort on the part of the author, who must respond to questions and suggestions. A provider that employs or has engaged a qualified researcher to conduct a significant study or program of studies should encourage the researcher to submit the report for this review. The report is often presented first at a research conference, which also has a review process, although usually less stringent. Afterwards, the researcher will work directly with the editorial board to revise and publish the report.

However, most research journals will not publish research on product or service impact, because they primarily publish theoretically or methodologically oriented research, where the question being addressed arises from a career program of research. In cases where research on product or service impact is published, often the details of implementation are removed in the journal editing, making the reports less useful to educators. While some research journals are devoted to program evaluation research, there are not enough editorial boards available to thoroughly review the many studies that are now beginning to emerge from the educational technology community.

There are three alternatives to traditional journal publication. First, a large number of conferences provide a professional audience for research results. These are often, but not always, a precursor to more formal publication. Examples include American Education Research Association (<http://www.aera.net>), American Evaluation

Association (<http://www.eval.org>) and Society for Research on Educational Effectiveness (<http://www.sree.org>).

Second, a wide range of repositories and websites exist from which reports can be disseminated. When using these, providers should begin by ensuring that their intervention evaluation studies are posted on their own website and on the websites of the participating research companies. Next, the reports should be posted on searchable repositories such as Education Resources Information Center or ERIC (<http://www.eric.ed.gov/>). 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.

Third, a version of peer review has also emerged in the form of government-funded organizations such as the What Works Clearinghouse (<http://ies.ed.gov/ncee/wwc/>) and the Best Evidence Encyclopedia (<http://www.bestevidence.org/>), both focused on educational programs, products and services. Many of the product effectiveness reports first become available through such sites. Unlike academic journals, these organizations actively seek research study reports applicable to the domains in which they are conducting reviews. And unlike academic journals, their review is more formulaic. Rather than engaging in a back-and-forth process for 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. Providers should recognize that not all research will meet the criteria of these organizations, but the criteria can be reviewed before a report is submitted.

**20. Make all formal evaluation research findings available upon request, regardless of the outcome, except in these instances: (a) a “failed experiment” where it is determined prior to review of outcomes data, for example, that the product or service was not implemented with fidelity, too few participants could be recruited, the study was too poorly designed or the data could not be collected; or (b) determination by the provider that the product or service must be improved and re-released, in which case the results can be considered as formative information for product improvement.**

An important goal of sharing research is to enhance trust between the provider community and consumers of education products and services. Scientific research builds 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. Providers should therefore make all research results available regardless of the outcome. 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. When all studies are reported regardless of outcome, we avoid what is called “publication bias,” where information is distorted through making results available contingent on the outcome.

At the same time, conducting research in schools is difficult, given the many practical challenges, which may include simply getting the product or service to be implemented. Recognizing these challenges, there are two important exceptions to this guideline’s suggestion that all research should be published regardless of the outcome.

First, an evaluation can be considered failed, and therefore not reported, in cases such as the following:

- 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.
- 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 three 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. If a flaw is discovered after the fact, the study should still be reported with a clear disclaimer about the limitation. In most cases, flaws in the experiment itself should not be grounds for withholding the results, although appropriate disclaimers should be made.

Second, a study can be considered formative and not reported if, in reviewing the results of the completed study and the intervention’s

general use, it is determined that significant changes and improvements are needed in the intervention before it can be successfully implemented. It is then assumed that the product will be significantly refined. In this case, the research can be considered a formative study. Because the report would be about a version of the intervention that would no longer be available to educators, releasing the results would be counterproductive and confusing.

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

### ***Accurately Translate Research for Customers***

This final set of guidelines 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 people 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 reports of the research do not overstate what has been established through rigorous analyses; otherwise, research and marketing claims will lose credibility over time.

#### **21. In the marketing literature for a product or service, accurately describe its impact – relative to the strength of the research design, quality of the evidence and size of the effect – using language that educators without research training can understand.**

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 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.

If the provider's internal research staff lack the qualifications, then the external researchers employed by the provider may be willing to assist in this task. However, where external researchers have completed the study, they may not be available or willing to assist the

provider or to review the correctness of the provider's translation. For example, submission of the full report may be the last of their contracted responsibilities. Moreover, they may consider helping to develop marketing materials somewhat of a conflict of interest. Providers should address these issues in the early stages. Either way, it is essential that this translation role be filled by a qualified and objective party.

When the marketing literature contains claims of causation, these claims should be substantiated by appropriate designs and methods of analysis that can eliminate plausible alternative explanations for its findings. Where the research has not fully eliminated other plausible explanations for observed achievement gains, it is important to be cautious about saying that the product or

service *caused* the gains. In such cases, it would be 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, whereas “The study found that our product had a significant impact on achievement levels” makes a stronger causal statement.

Claims referencing effect strength 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. It may also be

useful to translate the results into gains that are important from the viewpoint of the research site – say, percentage of students reaching proficiency in relation to a specific goal.

In the case of evidence of generalizability, several studies replicating a research finding provide stronger evidence than a single study. While standardized language for this dimension is not available, an indication should be provided as to the level of evidence available. For example, “In a study conducted in X school district, students gained 20 percentile points on the state test” does not make the claim that students will make 20-point gains in all contexts. In contrast, the statement that “Customers have consistently found impacts in the range of 20 percentile points” implies a much stronger level of evidence of generalizability to what potential new customers may find.

## **22. Cite the full research report any time the research or its findings is referenced.**

Any reference to specific research findings, or to product or service impacts based explicitly or implicitly upon evaluation research, should include a link to the full report(s) 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 the report to illustrate an effect but, without a reference back to the context of the research finding, the graph may be misleading.

## Conclusions

Educational technology providers engaging in research is not new. What is new is the relative attention now being paid by education decision makers to the research basis of product effectiveness. These Guidelines attempt to outline several practical considerations and best practices most unique to technology and to provider-sponsored research. Because the Guidelines are limited in scale and scope, they do not constitute a how-to manual, address all issues or cover issues in-depth. Instead, they are intended to be used to help providers ensure the quality of the evidence on their interventions available to customers.

We hope the Guidelines will help providers in understanding research as an ongoing process, rather than a one-time activity. This is especially important in light of 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 review and dissemination of research, while still an important avenue, are often not able to keep up with new versions of products or services.

In addition, because the educational technology industry is faced with a very

diverse marketplace, effectiveness research for their products and services cannot be conducted for every local curriculum and district population type. A reasonable sampling of contexts should be provided as educators look for a close match to their own criteria, all the while recognizing that an exact match may not exist. Alternatively, cooperative research between providers and customers can afford the local evidence that educators need to support their decisions. Such public-private partnerships are critical to meet education's needs. In fact, if the marketplace has to wait for providers to make localized evidence available before products and services can be implemented, innovation will be stalled as development cycles are forced to slow down to wait for evaluation research cycles to catch up.

The Guidelines provided here do not offer all the answers to how research can be financed or how it can provide timely answers for educators facing serious educational challenges and seeking effective solutions. The goal is to help providers understand the basic standards of research practice required for evidence that educators can use and to find productive and workable approaches to conducting and reporting that research.