Education Innovation Institute
WHITE PAPER

Bringing education to life.

UNIVERSITY of NORTHERN COLORADO

TAKING THE BOGEYMAN OUT OF EVALUATION PLANNING: RESOURCES FOR PRACTITIONERS

Evaluation has become an essential skill in the era of education reform. Major public and private funders generally require competitive grant applications to spell out plans for providing an independent, high-quality evaluation of a program's effectiveness, ideally coupled with recommendations for improving future outcomes. Rigorous evaluation also helps program administrators build a process of continuous improvement. The ongoing collection and analysis of data can tell whether a program is meeting goals and highlight areas in need of change or improvement. Correcting a problem early can save the expense of continuing an intervention that isn't working. And findings of an outcomes evaluation can contribute to the growing store of empirical knowledge about which programs work, why they work, and under what circumstances.

It's not uncommon for evaluations to seem daunting to educators who are at ease with any number of school programs but know little about conducting research. Practitioners also often wonder whether evaluations are the best use for the substantial time and money they require, resources that may seem better spent on the program itself. This paper discusses the many benefits of conducting a thorough, well-designed evaluation and seeks to demystify the process. It also stresses the importance of using rigorous evaluation methods.

Evaluations serve many purposes. They can assess whether interventions are bringing expected outcomes, weigh costs and benefits of a program, indicate whether a program is operating as planned, and tease out unintended consequences. It's common to think of evaluations as providing feedback for program stakeholders, but they can also provide guidance for anyone deciding whether to adopt a program or practice. Whatever your intended use, you will need certain knowledge and skills, including an understanding of the strengths and limitations of different types of evaluations and best practices for conducting them. For example, major foundations and government agencies increasingly require strong evidence that an intervention works before they will fund it; providing that evidence requires you to be able to assess the rigor and overall quality of existing evaluations.

There are several kinds of evaluations, each with different goals and implementation details. Summative evaluations of outcomes and/or impact seek to establish whether a program caused expected outcomes. Because they seek to establish causality these evaluations should be designed prior to program implementa-

UNIVERSITY of NORTHERN COLORADO
Education Innovation Institute

Bringing education to life.

tion using the most rigorous research methods possible. Resources to produce that rigor are a major focus of this Web site. We also note the importance of formative evaluations using data collected over the life of a project to assess whether a project is being implemented correctly and proceeding as expected. These evaluations, including performance monitoring and process evaluation, can lead to midcourse improvements to the program. One common type of evaluation, a cost benefit analysis, can be summative, formative or both, depending on when it is conducted.

Early, thorough planning is crucial

Whatever type of evaluation you intend to conduct, **planning should begin as soon as you start conceptualizing the program** to minimize operating constraints. Details such as program goals, the budget, the logic model for how you expect the program to work, and staffing are integral pieces of both the program and the evaluation. (For direction on developing a logic model, see the W.K. Kellogg Foundation Logic Model Development Guide.) Evaluation plans should determine analytic methods appropriate for the project, identify data sources, make provisions for regular data collection, and budget money and time to carry out the plans thoroughly. Also, the evaluation design must align with the purpose, scope, and needs of the intervention or program. For example, the evaluation of an outreach effort involving families and other stakeholders in a school would look very different from the evaluation of a new model for teaching math. Failure to fully anticipate evaluation needs as part of the original program design can result in serious obstacles at evaluation time, such as data unavailability or lack of stakeholder support. Such problems can result in wasted time and money and make it hard to produce clear evidence of whether a program worked.

The U.S. Department of Health and Human Service's Program Manager's Guide To Evaluation considers this early work essential. Its advice includes:

> **Invest heavily in planning**. Invest both time and effort in deciding what you want to learn from your evaluation. This is the single most important step you will take in this process. Consider what you would like to discover about your program and its impact on participants, and use this information to guide your evaluation planning.

> **Integrate the evaluation into ongoing activities of the program**. Program managers often view evaluation as something that an outsider "does to" a program after it is over, or as an activity "tacked on" merely to please funders. Unfortunately, many programs are evaluated in this way. This approach greatly limits the benefits that program managers and staff can gain from an evaluation. Planning the evaluation should begin at the same time as planning the program so that you can use evaluation feedback to inform program operations (p. 3).

The definition of evaluation in the U.S. General Accountability Office's Designing Evaluations, 2012 Revision notes the importance of evaluation in continuous improvement:

UNIVERSITY of
NORTHERN COLORADO
Education Innovation Institute

Bringing education to life.

> A program evaluation is a systematic study using research methods to collect and analyze data to assess how well a program is working and why. Evaluations answer specific questions about program performance and may focus on assessing program operations or results. Evaluation results may be used to assess a program's effectiveness, identify how to improve performance, or guide resource allocation (p. 3).

The GAO's terms "systematic study" and "specific questions" are key to the planning process. Evaluations are not haphazard; the framework stems organically from the program's goals, theoretical underpinnings, design, methods, and logic model. Similarly, the questions addressed by an evaluation must be tightly bound to the program's core components. This inter-connectedness is why experts advise choosing – and vetting – the person to conduct the external evaluation early in the planning process to make sure her skills and capacity match your needs and to get her advice on evaluation design and data collections that should begin early.

Chapter 1 of the Handbook of Practical Program Evaluation, Planning and Designing Useful Evaluations, explains some reasons why early planning is important and describes essential steps to be taken early.

> Desired data can be more readily obtained if provision is made for data collection from the start of the program, particularly for such information as clients' preprogram attitudes and experiences. These sorts of data might be very difficult, if not impossible, to obtain later.
>
> Planning an evaluation project requires selecting the measures that should be used, an evaluation design, and the methods of data collection and data analysis that will best meet information needs. Evaluators should be able to anticipate how the evaluation results might be used and how decision making might be shaped by the availability of the performance data collected. However, it is important to recognize that evaluation plans are organic and likely to evolve. Figure [1, reproduced on the next page] displays the key steps in planning and conducting an evaluation. It highlights many feedback loops in order to stress how important it is for evaluators to be responsive to changes in context, data availability, and their own evolving understanding of context (pp. 17-18).

The Handbook also notes that the greatest benefits come when evaluations are approached as more than a one-shot thumbs-up or -down. Information gathered systematically throughout a project can identify strengths and weaknesses of the program -- and the evaluation plans -- ideally leading to improvements in both. Figure 1 illustrates the formative processes— a series of feedback loops – through which ongoing data gathering can suggest ways to refine the questions and methods of an evaluation. Similar ongoing feedback and data analyses can lead to midcourse corrections to a program and improve delivery.

UNIVERSITY of
NORTHERN COLORADO
Education Innovation Institute

Bringing
education
to life.

Figure 1. Feedback loops for using ongoing data collection to refine evaluations
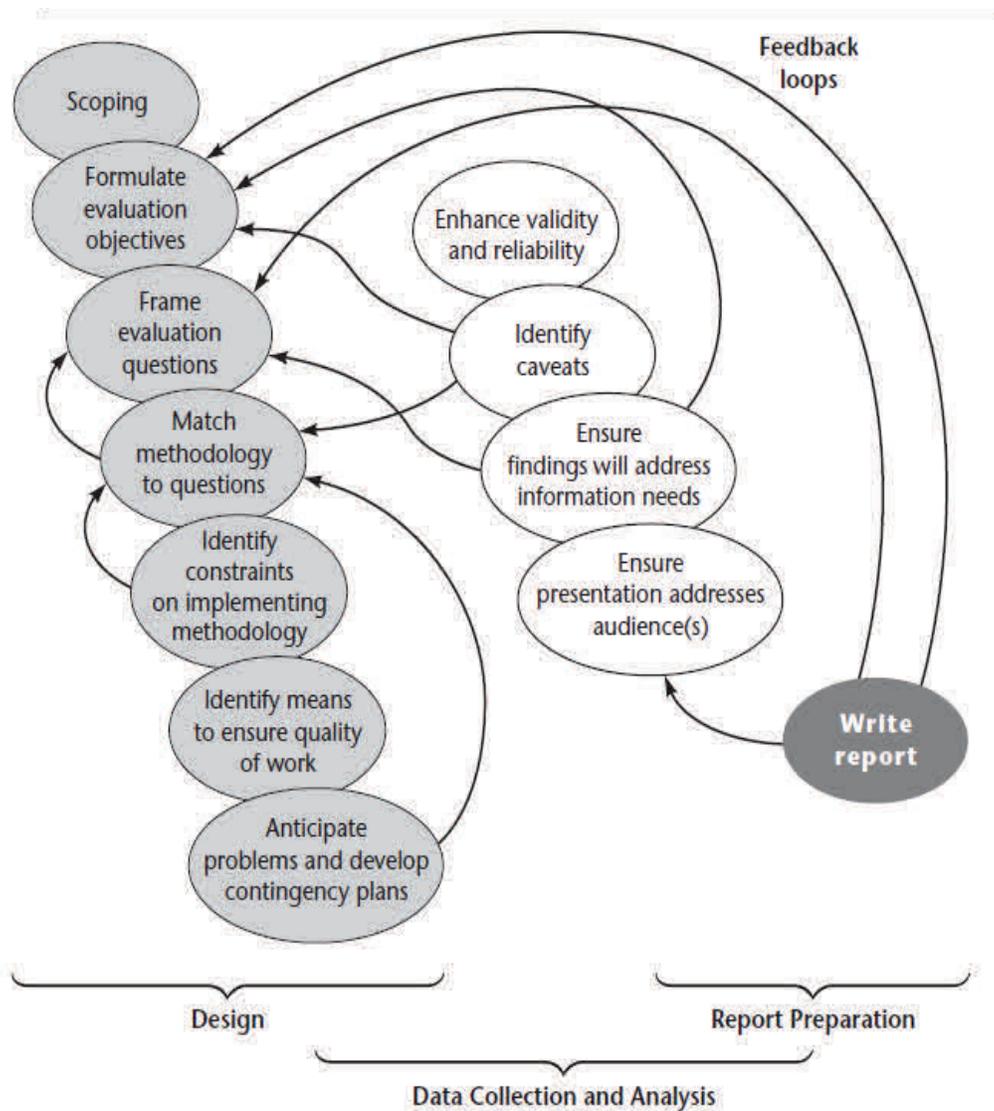
Source: Planning and Designing Useful Evaluations, in Wholey, J.S., Hatry, H.P., and Newcomer, K.E., Eds., *Handbook of Practical Program Evaluation, 3rd Edition* (2010). Jossey-Bass. Used with permission from the publisher.

The Need for Rigorous Evaluation Methods

The federal government has promoted the use of impact evaluations in its pursuit of "evidence-based policy" to direct more public money toward programs with a demonstrated likelihood of success. In 2005, the U.S. Department of Education announced in the Federal Register that it would "focus Federal financial assistance on expanding the number of programs and projects Department-wide that are evaluated under rigorous scientifically based research methods." Since then, the DOE has given strong preference to

UNIVERSITY of
NORTHERN COLORADO
Education Innovation Institute

Bringing
education
to life.

the use of experimental or quasi-experimental designs to provide evidence of whether an intervention or program affects desired outcomes, often improved student achievement.

In Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide the DOE explains its preference for rigorous evaluation designs as a way to bring to education the evidence driven progress experienced by medicine and a few other fields.

> The federal *No Child Left Behind Act of 2001*, and many federal K-12 grant programs, call on educational practitioners to use "scientifically-based research" to guide their decisions about which interventions to implement. As discussed below, we believe this approach can produce major advances in the effectiveness of American education. Yet many practitioners have not been given the tools to distinguish interventions supported by scientifically-rigorous evidence from those which are not (p. iii).

> Life and health in America has been profoundly improved over the past 50 years by the use of medical practices demonstrated effective in randomized controlled trials. These research-proven practices include: (i) vaccines for polio, measles, and hepatitis B; (ii) interventions for hypertension and high cholesterol, which have helped bring about a decrease in coronary heart disease and stroke by more than 50 percent over the past half-century; and (iii) cancer treatments that have dramatically improved survival rates from leukemia, Hodgkin's disease, and many other types of cancer.

> Similarly, welfare policy, which since the mid-1990s has been remarkably successful in moving people from welfare into the workforce, has been guided to a large extent by scientifically-valid knowledge about "what works" generated in randomized controlled trials (p. iv).

The great benefit of randomized controlled trials (RCTs) is that random assignment of participants to the treatment group (those in the program) or the control group removes the problem of selection bias so that any difference in outcomes between the two groups can be attributed to the program as causal. While education practitioners and researchers recognized the value of such solid evidence they also knew from experience that experiments – in this context, studies using RCTs – can be very hard to conduct in schools for ethical, financial, practical, and, at times, political reasons. The American Educational Research Association and the American Evaluation Association noted these difficulties in correspondence with the DOE and urged the department to give equal weight to well-executed evaluations that are aligned to the program design. The AEA particularly noted that in many situations "the evaluation ought to include interviews, observations, case studies, surveys and other strategies to understand causality and to provide the information needed to improve the educational experience."

UNIVERSITY *of*
NORTHERN COLORADO
Education Innovation Institute

Bringing education to life.

Several publications in our Resources for Practitioners section explain what makes research rigorous and describe different kinds of experimental and quasi-experimental techniques that are most commonly considered rigorous. Quasi-experimental designs use statistical methods to come as close as possible to the benefits of an RCT without having to use random assignment by creating a comparison group with salient characteristics that are closely matched with those of program participants. While quasi-experimental methods do not completely eliminate selection bias or create a true counterfactual, they can come close enough to make strong inferences about a program's effectiveness.

Become a Skilled Evaluator of Evaluations

Rigorous evaluation methods don't ensure that results will be unequivocal, however. As the National Forum on Early Childhood Program Evaluation notes in its Early Childhood Program Evaluations: A Decision-Maker's Guide,"This dilemma is illustrated by the intense debate that often ensues among dueling experts who reach different conclusions from the same data about whether a program is effective or whether its impacts are large enough to warrant a significant investment of public and/or private funds." The guide then provides advice on how to ignore such politicized debates by assessing and interpreting the contents of an evaluation objectively. The guidance is broken into five broad categories, each containing readable explanations that are worth checking out:

- Is the evaluation design strong enough to produce trustworthy evidence?

- What program services were actually received by participating children and families and comparison groups?

- How much impact did the program have?

- Do the program's benefits exceed its costs?

- How similar are the programs, children, and families in the study to those in your constituency or community?

One place to learn more about rigorous research designs is the DOE's What Works Clearinghouse, which contains a small but growing number of studies that meet the DOE's standards for rigor. There you can see how the federal government uses empirical evidence to declare that educational interventions work – or don't. You can also see how various statistical methods are applied and read critiques by experts about the strengths and weaknesses in the design and execution of each study.

UNIVERSITY *of*
NORTHERN COLORADO
Education Innovation Institute

Bringing education to life.