Developing An Effective Evaluation Strategy: Suggestions For Federal and State Education Officials

Coalition for Evidence-Based Policy
A Project Sponsored by The Council for Excellence in Government

February 2007

This publication was produced by the Coalition for Evidence-Based Policy, with funding support from the William T. Grant Foundation, Edna McConnell Clark Foundation, and Jerry Lee Foundation. An earlier version of this document was prepared for U.S. Office of Management and Budget (OMB), and circulated by OMB to federal math and science education programs as part of the Congressionally-established Academic Competitiveness Council.

We wish to acknowledge the valuable comments we received, in preparing this document, from Howard Rolston, Robert Slavin, and Ricky Takai.

We welcome comments and suggestions on this document (jbaron@excelgov.org).

Purpose of this Document:

To assist federal and state education programs in developing an effective evaluation strategy – one which is likely to succeed in identifying, in rigorous evaluations, at least a few interventions within the program that produce meaningful improvements in the educational outcomes.

Such research-proven interventions, if disseminated through the program, may then offer the key to important improvements in the program's overall effectiveness.

In this document, the term "intervention" refers generically to any project, practice, or strategy funded by a federal or state education program. The term "rigorous evaluation" refers to studies that are capable of producing valid evidence of an intervention's true effect, such as well-designed randomized controlled trials or, if not feasible, well-matched comparison-group studies.[1]

This document focuses on a few key principles to consider in developing an evaluation strategy. It is not intended as a comprehensive listing of all elements comprising a successful strategy.

Suggestion 1:

Focus your program's rigorous evaluations on those interventions which are most promising, as opposed to trying to evaluate everything.

Our rationale is that a strategy focused on the most promising interventions stands the greatest chance of producing a few valid findings of intervention effectiveness that you can then apply more broadly to improve the program's overall performance. A less-focused strategy, by contrast, may well produce only a series of negative findings, for the following reason. When rigorous evaluations are carried out in education and almost every other field, they typically find that many interventions are ineffective or marginally effective. Those found to produce meaningful positive effects tend to be the exception – sometimes the rare exception. Thus, if you do not carefully focus your rigorous evaluations on the most promising candidates, you are likely to miss the few opportunities to produce a valid finding of intervention effectiveness. This is especially likely to be true if your program's resources for evaluation are fairly limited, allowing you to sponsor only a small number of rigorous evaluations.

A good place to start in identifying highly-promising interventions is the literature of previous research – including, for example, suggestive findings from less-rigorous studies (such as a partially-matched comparison-group study) that an intervention is highly effective. Or sometimes a program provider or educator will develop what seems like a great idea that merits rigorous evaluation even in the absence of such earlier research findings.

[1] For an overview of what constitutes a rigorous evaluation, see Which Study Designs Can Produce Rigorous Evidence of Program Effectiveness? A Brief Overview, Coalition for Evidence-Based Policy working paper, January 2006.

Suggestion 2:

You may also wish to focus your rigorous evaluations on interventions that your program is heavily investing in, so as to produce evidence you can use to decide whether to continue that investment.

For example, if many of your program grantees are adopting a particular educational curriculum or teacher professional development model, you may wish to rigorously evaluate that curriculum or model so as to determine whether it works and should be continued.

Suggestion 3:

Make sure that the type of evaluation used to test an intervention is appropriate for the stage of that intervention's development (e.g., early-stage development/refinement, versus later-stage scale-up of a mature intervention).

When an intervention is still in the early stages of development and refinement, it often makes sense to test the intervention, and/or various components of it, in small-scale, low-cost evaluations. In some cases, these evaluations might use more rigorous designs (e.g., small-scale randomized controlled trials); in other cases, they might use less rigorous designs (e.g., matched or partially-matched comparison-group studies).

As an illustrative example, one might evaluate whether the addition of a new computer-assisted tutoring intervention to a typical second-grade math curriculum improves math achievement by randomly assigning perhaps 100 second-grade students in 1-2 schools to (i) the existing math curriculum, or (ii) the existing curriculum plus the computer-assisted tutoring. (Alternatively, one could do a matched or partially-matched comparison-group study if random assignment is not feasible.) The evaluation would compare math achievement outcomes for the two groups on school-administered standardized tests six months later. If this evaluation finds that the computer-assisted tutoring has little or no effect, it might be modified and evaluated again in such a small-scale study.

If the intervention proves effective, however, a next stage might be to evaluate it in a larger study that uses random assignment (if at all feasible), is conducted across several schools or districts where this intervention would typical be deployed, and measures math achievement outcomes over a period of 2-3 years. Depending on how the computer-assisted tutoring is administered (within or outside the students' regular classroom), the evaluation may need to randomly assign classrooms and teachers – or even whole schools – to intervention and control groups.

If found effective in a rigorous evaluation of this type, the intervention would presumably be ready for wider dissemination. If resources allow, one might build a large, rigorous evaluation into this dissemination process, to make sure the intervention continues to be effective as it is scaled up.

Suggestion 4:

Recognize that it is often possible to conduct a well-designed randomized controlled trial at modest cost by measuring outcomes using administrative data that are already collected for other purposes.[2]

Such data can reduce a trial's cost by eliminating what is typically the most labor-intensive and costly part of a trial – namely, locating the individual sample members at various points in time after the intervention is completed, and administering surveys, tests, interviews, and/or observations to obtain their outcome data. If, instead, key outcome data are readily available for most or all sample members from administrative data sources, the cost of data collection can be reduced to a nominal amount. This can be true even for trials involving large numbers of students, including trials that randomize whole schools or classrooms, rather than individuals. The cost of a trial may be as low as $50,000 - $100,000 under some conditions.

Illustrative examples of administrative data that might be used to measure outcomes in a low-cost randomized controlled trial of an educational intervention at the K-12 or post-secondary level include: scores on standardized achievement tests, disciplinary suspensions, attendance, grade retentions, special education placements, attendance, high school graduation, employment and earnings after graduation, criminal arrests, enrollment in and/or completion of particular courses (e.g., advanced math or science), attainment of an undergraduate or graduate degree, and research publications and citations.

Suggestion 5:

Provide clear guidance and/or technical assistance to researchers conducting rigorous evaluations, to prevent common flaws in study design and implementation.

Such guidance might include a short checklist, in the program solicitation, of key parameters that the evaluation is expected to adhere to, in order to ensure its validity.[3] The technical assistance might consist of brief consulting by phone and email with an expert in rigorous evaluations, to identify any serious flaws in the study's design before it gets underway, and to help with problem-solving advice during the course of the study. The reason for these suggestions is that many, perhaps most, randomized controlled trials in education contain design and implementation flaws that may undermine the validity of the study's findings – flaws which, in many cases, could readily have been prevented.

Illustrative examples of common flaws include: (i) using a sample that is too small to detect meaningful effects of the intervention; (ii) randomizing groups (such as classrooms or schools) but conducting statistical analysis as if individual students had been randomized; (iii) failing to collect and analyze outcome data for those members of the intervention group that don't participate in or

[2] For additional information, see When Is It Possible To Conduct a Randomized Controlled Trial in Education at Reduced Cost, Using Existing Data Sources? A Brief Overview, Coalition for Evidence-Based Policy, February 2007.

[3] Such a checklist can be found in Key Items To Get Right When Conducting A Randomized Controlled Trial in Education, Coalition for Evidence-Based Policy and What Works Clearinghouse, 2005.

complete the intervention; and (iv) failing to obtain outcome data for the vast majority of sample members originally randomized.

Suggestion 6: Where To Start?

If your program is new to rigorous evaluations, you might start by trying to get a small number of such evaluations (e.g., one or two) underway, to evaluate highly-promising intervention(s) in your program.

The rationale for this suggestion is that it takes time for program officials, funding awardees, and evaluators to develop the experience needed to advance a sizeable number of rigorous evaluations that will produce a stream of important findings about "what works" within the program. A first step in gaining that experience is to sponsor one or a few high-quality evaluations, perhaps in close consultation with a research agency or other organization that has significant experience with such evaluations. Such a first step – in addition to providing the needed learning experience – will hopefully demonstrate to key agency decision-makers that rigorous evaluations are indeed practical within your program, and that they can produce valuable knowledge which the program can use to improve its overall performance.