This guide was authored by the nonprofit, nonpartisan Coalition for Evidence-Based Policy. The Coalition wound down its operations in 2015, and the group's leadership and core elements of its work have been integrated into Arnold Ventures (a philanthropic organization), as Arnold's Evidence-Based Policy initiative. The initiative is making this and other Coalition documents available as a public resource.

Practical Evaluation Strategies for Building a Body of Proven-Effective Social Programs

Suggestions for Research and Program Funders

Coalition for Evidence-Based Policy

A NONPROFIT, NONPARTISAN ORGANIZATION

October 2013

This publication was produced by the nonprofit, nonpartisan Coalition for Evidence-Based Policy, with funding support from the Hewlett Foundation, the William T. Grant Foundation, and the MacArthur Foundation.

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

Purpose and Background

Purpose: To recommend evaluation strategies that can efficiently build a body of social programs, practices, and strategies ("interventions") backed by credible evidence of effectiveness.

Target audience: Government agencies, philanthropic foundations, and other organizations that fund research studies and/or program services. We believe that program developers and providers, while not a primary audience for this brief, may also find its suggestions useful in building evidence about their program's effectiveness.

Background: This brief uses, as a framework for its recommendations, the Common Evidence Guidelines of the Institute of Education Sciences (IES) and National Science Foundation (NSF).[1] The recently-developed Guidelines "identify the spectrum of study types that contribute to development and testing of interventions and strategies, and … specify expectations for the contributions of each type of study." Although the Guidelines focus on education research, we believe their core concepts are also applicable to other fields of social policy research and evaluation.

What this brief adds: Concrete suggestions for using evaluations to identify interventions that produce sizable effects on important life outcomes, recognizing that most interventions produce weak or no effects.

A. **Most interventions, when rigorously evaluated, are found to produce weak or no effects compared to services-as-usual.** This pattern occurs not just in social spending but in other fields, and applies to many interventions widely believed to be effective, based on less-rigorous studies and/or expert opinion. The following examples – findings from randomized controlled trials (RCTs), which are generally considered the strongest method of evaluating effectiveness – illustrate the pattern:

   - **Education:** Of the 90 interventions evaluated in RCTs commissioned by IES since 2002, approximately 90% were found to have weak or no positive effects.[2]

   - **Employment/training:** Of the 13 interventions evaluated in Department of Labor RCTs that have reported results since 1992, about 75% were found to have found weak or no positive effects.[3]

   - **Medicine:** Reviews have found that 50-80% of positive results in initial ("phase II") clinical studies are overturned in subsequent, more definitive RCTs ("phase III").[4]

   - **Business:** Of 13,000 RCTs of new products/strategies conducted by Google and Microsoft in recent years, 80-90% have reportedly found no significant effects.[5]

B. **However, some interventions have been found highly effective in rigorous evaluations, such as:**

   - **Nurse-Family Partnership** – a nurse visitation program for low-income, first-time mothers during pregnancy and children's infancy (shown in rigorous RCTs to reduce child maltreatment by 20-50% and, for the most at-risk children, to improve educational outcomes – e.g., 8% higher GPA).[6]

   - **H&R Block College Financial Aid Application Assistance** – streamlined personal assistance for low and moderate income families with a dependent child near college age (shown in a large multisite RCT to increase college enrollment and persistence by 29%, over a 3½-4 year period).[7]

C. **Our recommendations seek to maximize the chance of finding such interventions producing sizable, sustained improvement in people's lives.** Doing so, we suggest, is the main path through which evaluation activities can improve the human condition.

Recommendation 1

Recognize that many interventions are in the early (design and development) stage, best suited for feasibility testing and further refinement, and that it may be premature to evaluate their impact on their ultimate targeted outcomes (e.g., teen pregnancy rates).

A. To be considered ready for an impact evaluation, an intervention generally should be able to show that it –

(i) Can be successfully implemented, with close adherence to its key elements, in the settings where it would normally be delivered (e.g., public schools, community health clinics); and

(ii) Has persuasive logic or initial evidence to suggest it may have an impact on its targeted outcomes – as discussed further in recommendations 3 and 4.

B. Funding impact evaluations of interventions that do not yet meet the above conditions would likely be an inefficient use of resources, producing findings of weak or no impact for interventions that need further development to be effective, or might never be able to establish initial evidence of promise.

C. For such interventions, funders might do better to focus on moving the interventions closer to meeting the above conditions, through activities such as the following. These activities fall within the "Design and Development" stage described in the IES/NSF Guidelines.

- **Conducting feasibility studies to establish whether the intervention can be successfully implemented** – e.g., can enroll and retain participants, and deliver the intervention in adherence to a well-defined protocol, in settings where it would typically be delivered;
- **Assessing whether program participants understand, and begin to apply, the skills taught by the intervention**;
- **Using results of the above activities to refine and improve the intervention**;
- **Developing a manual documenting the content of the intervention**, and a training and supervision program for those who will deliver it;
- **Creating a data system to support successful implementation of the intervention** – i.e., to track whether its key elements are being faithfully delivered, and identify any substantial deviations that need correcting or indicate a need to refine the intervention; and
- **Creating or identifying a data system to support an impact evaluation (should one eventually be undertaken)** – i.e., a system to track outcomes of program participants as well as outcomes of a control or comparison group of nonparticipants.

Recommendation 2

For interventions that are ready for an impact evaluation, use scientifically-rigorous methods to measure impact – wherever feasible, RCTs – and select evaluators with demonstrated success conducting such studies.

A. The goal is to ensure the evaluation produces a true measure of the intervention's effect, and doesn't consume time and resources on methods that may not yield a credible result.

B. Well-conducted RCTs are the most credible method of evaluating impact, per the IES/NSF Guidelines, as well as evidence standards articulated by the National Academy of Sciences,[8] Congressional Budget Office,[9] U.S. Preventive Services Task Force,[10] Food and Drug Administration,[11] and other respected scientific bodies.

Specifically, the IES/NSF Guidelines recommend: "Generally and when feasible, [impact studies] should use designs in which the treatment and comparison groups are randomly assigned," – i.e., RCTs. The Guidelines intend this recommendation to apply to the full range of impact evaluations, from early-stage "efficacy" studies to later-stage "scale-up" evaluations. Similarly, a National Academy of Sciences report recommends that evidence of effectiveness generally "cannot be considered definitive" without ultimate confirmation in well-conducted RCTs, "even if based on the next strongest designs."[8]

C. When an RCT is not feasible –

- **The IES/NSF Guidelines state that "quasi-experimental designs, such as matched comparison groups or regression discontinuity designs, are acceptable only when there is direct compelling evidence demonstrating the implausibility of common threats to internal validity."** The phrase "threats to internal validity" means study features that could produce incorrect estimates the intervention's impact in the study sample.[12]

- **We have published a brief that summarizes which non-randomized study designs are most likely to avoid such threats and thus produce valid estimates of impact**.[13]

D. We recommend selecting an evaluator that has previously conducted a high-quality RCT (or other rigorous non-randomized study if an RCT is not feasible). For simplicity of discussion, the following assumes an RCT will be used.

- **We suggest this because many attempted RCTs fail to yield credible impact findings because of key flaws in study design or implementation that could have been avoided.** Examples of common flaws include: (i) using a sample too small to detect meaningful effects of the intervention; (ii) failing to obtain and analyze outcome data for a high proportion of the original sample; and (iii) measuring intermediate outcomes that lack practical and policy importance (e.g., student-reported expectations to stay in school, rather than actual dropout and graduation rates).

- **We believe it is essential to select an evaluator that is aware of such pitfalls, and has demonstrated the ability to avoid them in the evaluator's prior studies.**

- **Thus we suggest the following, streamlined process for vetting potential evaluators**:

  (i) Ask evaluators to provide one or two prior RCTs that they have conducted; and

  (ii) Have an independent research expert conduct a brief, top-level review of the RCTs to see if they were well-conducted and produced credible findings – perhaps using an established checklist for reviewing such studies.[14]

Recommendation 3

Wherever feasible, embed low-cost rigorous evaluations – preferably RCTs – in program initiatives that governmental or philanthropic organizations are implementing anyway as part of their usual (non-research) program operations.

A. Recently, researchers have shown it is often possible to conduct RCTs at low cost and burden by –

(i) Embedding random assignment in social program initiatives that are being implemented anyway. Government and foundations fund a vast array of strategies and approaches and, over time, new initiatives and reforms are often launched. Credible evaluations can be embedded in many of these efforts – for example, by (i) using a lottery process – i.e., random assignment – to determine who will be offered program services (since programs often do not have sufficient funds to serve everyone who is eligible); or (ii) randomly assigning some individuals to the program's usual approach (e.g., transitional jobs for ex-offenders) versus a revised model that is being piloted (e.g., transitional jobs plus drug treatment), to see if the new model produces better outcomes.

– and –

(ii) Using administrative data that are collected already for other purposes to measure the key outcomes, rather than engaging in original – and often costly – data collection (e.g., researcher-administered interviews, observations, or tests). In many jurisdictions, administrative data of reasonable quality are available to measure outcomes such as child maltreatment rates, employment and earnings, student test scores, criminal arrests, receipt of government assistance, and health care expenditures.

B. Such piggybacking on ongoing efforts/resources enables many more RCTs to go forward, by reducing their cost as much as tenfold. We recently published a short brief containing examples of well-conducted, low-cost RCTs in diverse policy areas that cost between $50,000 and $300,000.[15]

C. Illustrative example: A large, low-cost RCT ($50,000) was embedded in New York City's $75 million Teacher Incentive Program, generating definitive evidence about the program's impact. The program, launched in 2008, provided low-performing public schools that increased student achievement and other key outcomes with an annual bonus, to be distributed to teachers. Because the city did not have sufficient funding to provide the program to every low-performing school in New York, the city allowed researchers to use a lottery – i.e., random assignment – to determine which of 396 eligible schools would receive the program (the treatment group) and which would not (the control group). The researchers then measured student outcomes in both groups using administrative data (e.g., state test scores) that the school district already collected for other purposes.

Because random assignment was baked into the design of the program, and all outcomes were measured with administrative data, the study was very inexpensive – $50,000. Yet it produced a definitive result: over a three-year period, the program had no effect on student achievement, attendance, graduation rates, behavior, GPA, or other outcomes (versus control schools). Based in part on these results, the city ended the program, freeing up resources for other efforts to improve student outcomes.[16]

D. Because of the low cost, we recommend embedding such studies in any initiative where conditions allow (e.g., lottery and administrative data can be used), and that is at least minimally promising. Minimal promise might consist of evidence that the initiative is reasonably well-implemented and has a plausible rationale for how it might improve participant outcomes.

E. The end goal is to greatly increase the number of interventions rigorously tested, so as to more rapidly grow the number proven to work. The basic concept was captured by IBM pioneer Thomas Watson more than half a century ago: "If you want to succeed, double your failure rate."

Recommendation 4

When considering investments in more comprehensive and/or costly rigorous evaluations, be highly selective about <u>what</u> to evaluate, making sure an intervention's initial evidence is sufficiently promising to merit such a study.

A. As traditionally carried out, rigorous impact evaluations – both RCTs and high-quality non-randomized studies – are often specially-crafted projects costing several million dollars, in which evaluation funds pay for such items as delivery of the intervention, recruitment of a sample population, site visits, implementation research, and data collection through researcher-administered interviews, observations, or tests. The study cost can vary greatly, depending on how many of the above items will be funded, as well as factors such as sample size, number of study sites, and duration of the study. Cost typically exceeds $1 million and in some cases may reach tens of millions, with the largest cost components usually being data collection and intervention delivery.

B. Such studies, when focused on highly-promising interventions, can generate evidence that is of great value for decision-making about whether and how to scale up the intervention. Such evidence may include, for example: (i) corroboration of earlier impact findings (e.g., from a low-cost RCT) in a different sample and setting, thus generating strong, replicated evidence of effectiveness; (ii) estimates of the intervention's effect on outcomes other than those measurable with administrative data; (iii) the subgroups and conditions in which the intervention is most effective; (iv) detailed information on the services received by intervention participants, and how they differ from any services received by the control group; (v) possible reasons why the intervention produced its effect; and (vi) how the intervention's benefits compare to its costs.

C. But if focused on less-promising interventions, such studies can be a costly and inefficient use of evaluation funds, because of the high likelihood of finding no meaningful impacts, as has been the clear pattern in past rigorous evaluations (as described earlier). Too often, in these cases, large investments are made in extensive implementation research and data collection (e.g., to learn about implementation challenges, mechanisms through which impacts occur, and effects on a broad set of outcome measures) for interventions that, in the end, are not found sufficiently effective to warrant adoption.

D. Thus, we recommend setting a fairly high bar for evidence of promise before sponsoring a comprehensive/costly evaluation – and higher still for the most expensive studies. Specifically:

(i) As a threshold condition, we suggest requiring evidence that the intervention can be successfully implemented, and has a persuasive rationale, as discussed in recommendation 1.

(ii) In addition, we suggest carefully reviewing prior studies of the intervention (perhaps in consultation with a research expert) to make sure it has highly-promising evidence, such as:

- **Prior rigorous evaluations that show sizable impacts for the intervention, but are not yet conclusive** – for example, because the earlier evaluations only measured short-term impacts; used closely-matched comparison groups but not random assignment; consisted of a single-site RCT that needs replication to establish strong evidence; or measured impacts using low-cost administrative data (e.g., hospitalization rates) that need corroboration with researcher-administered measures (e.g., of participant health and functional ability).

- **Prior rigorous evaluations that have established the intervention's effectiveness in a different population than that proposed for the current study** (e.g., rural whites, versus urban African Americans);

- **Prior rigorous evaluations that have established the effectiveness of a closely-related intervention** (e.g. a high-cost version of the program delivered by specialized personnel, versus the lower-cost version that the current study seeks to evaluate);

– and/or –

- **Prior evaluations that were not fully rigorous, but consistently show large impacts across a diverse range of methods, researchers, and settings.**

Recommendation 5

Fund each comprehensive and/or costly evaluation in seamless stages that first measure the intervention's impact on the main outcome(s) of interest, and add more comprehensive data collection and analysis only if such impact is found.

A. This approach, developed by an expert panel for the Department of Labor's Employment and Training Administration,[17] i**s designed (in the panel's words) to –**

- **"improve efficiency by allowing mid-course adjustments in the evaluation plans** (including, in some cases, reducing scope to omit portions of study plans that are not likely to yield credible and useful evidence);" and

- **"increase the number of interventions that can be rigorously tested within [a given] evaluation budget."**

B. Following the panel's recommendation, we specifically suggest –

(i) A first stage that uses a rigorous evaluation (preferably RCT) to measure the intervention's impact on the primary outcome(s) of interest – if possible using low-cost administrative data. In this stage, we also suggest that the study obtain basic data on the operation and cost of the intervention, to provide a general understanding of what is being evaluated. This might include, for example: (a) the number of individuals who participated, how many hours they participated, and what skills were taught; and (b) the intervention's cost, obtained by aggregating the cost of labor, materials, and other resources used in its delivery.

– and –

(ii) Subsequent stages involving more comprehensive data collection and analysis, which would go forward only if warranted by positive impact findings in the first stage. Examples of activities that may be useful in these stages include (a) implementation research, conducted retrospectively in the sites that produced the largest impacts and those that produced the smallest impacts, to identify factors that might account for the different-sized impacts; and (b) a second impact evaluation (e.g., in a new cohort of sample members) designed to confirm, through replication, that the intervention is effective, and also to identify reasons why the intervention produced its effect, the conditions and subgroups in which it is most effective, and its impacts on a broader set of outcome measures.

Appendix A to the panel's recommendation contains an example of a major Department of Labor-sponsored RCT that used this staged approach to build valuable evidence in a cost-effective manner about the effectiveness of a Department program for unemployed workers.

Conclusion:

A strategic approach to evaluation, which uses rigorous methods and takes account of the challenges in finding truly effective interventions, can efficiently build the body of proven approaches to improving people's lives.

References

[1] Institute of Education Sciences (of the U.S. Department of Education) and National Science Foundation, *Common Guidelines for Education Research and Development*, August 2013, linked here.

[2] Coalition for Evidence-Based Policy, *Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects*, July 2013, linked here.

[3] This is based on a count of results from the Department of Labor RCTs that reported results between 1992 and 2013, as identified through the Department's research database (link).

[4] John P. A. Ioannidis, "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, vol. 294, no. 2, July 13, 2005, pp. 218-228. Mohammad I. Zia, Lillian L. Siu, Greg R. Pond, and Eric X. Chen, "Comparison of Outcomes of Phase II Studies and Subsequent Randomized Control Studies Using Identical Chemotherapeutic Regimens," *Journal of Clinical Oncology*, vol. 23, no. 28, October 1, 2005, pp. 6982-6991. John K. Chan et. al., "Analysis of Phase II Studies on Targeted Agents and Subsequent Phase III Trials: What Are the Predictors for Success," *Journal of Clinical Oncology*, vol. 26, no. 9, March 20, 2008.

[5] Jim Manzi, *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*, Perseus Books Group, New York, 2012, page 128 and 142. Jim Manzi, *Science, Knowledge, and Freedom*, presentation at Harvard University's Program on Constitutional Government, December 2012, linked here.

[6] A summary of the evidence on NFP, including citations to the original study reports, is linked here.

[7] Eric P. Bettinger, Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment," *Quarterly Journal of Economics*, August 2012, vol. 127, no. 3, pp. 1205-1242.

[8] National Research Council and Institute of Medicine, *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities*, Mary Ellen O'Connell, Thomas Boat, and Kenneth E. Warner, Editors (Washington DC: National Academies Press, 2009), recommendation 12-4, p. 371, linked here.

[9] *CBO's Use of Evidence in Analysis of Budget and Economic Policies*, Jeffrey R. Kling, Associate Director for Economic Analysis, November 3, 2011, page 31, linked here.

[10] U.S. Preventive Services Task Force, "Current Methods of the U.S. Preventive Services Task Force: A Review of the Process," *American Journal of Preventive Medicine*, vol. 20, no. 3 (supplement), April 2001, pp. 21-35.

[11] The Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.126, linked here.

[12] The IES/NSF Guidelines go on to say that such threats to internal validity "might include selection bias in the case of matched comparison groups, or, in the case of regression discontinuity designs, nonlinearities of treatment impacts over a policy relevant span around the 'cut point.'" See reference 1, p. 21.

[13] Coalition for Evidence-Based Policy, *Which Comparison-Group ("Quasi-Experimental") Study Designs are Most Likely to Produce Valid Estimates of a Program's Impact?: A Brief Overview and Sample Review Form*, February 2012, linked here.

[14] An example of such a checklist is Coalition for Evidence-Based Policy, *Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence*, February 2010, linked here.

[15] Coalition for Evidence-Based Policy, *Rigorous Program Evaluations on a Budget: How Low-Cost Randomized Controlled Trials Are Possible in Many Areas of Social Policy*, March 2012, linked here.

[16] Roland G. Fryer, "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," NBER Working Paper No. 16850, March 2011, linked here.

[17] Rebecca Maynard, Larry Orr, and Jon Baron, *Increasing the Success of Evaluation Studies in Building a Body of Effective, Evidence-Based Programs: Recommendations of a Peer-Review Panel*, prepared for the Employment and Training Administration of the U.S. Department of Labor, June 2013, linked here.