Which Study Designs Are Capable of Producing Valid Evidence About A Program's Effectiveness?

A Brief Overview

Coalition for Evidence-Based Policy
A NONPROFIT, NONPARTISAN ORGANIZATION

October 2014

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

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

Which Study Designs Are Capable of Producing Valid Evidence About A Program's Effectiveness?

This guide is addressed to policy officials, program providers, and researchers who are seeking to (i) identify and implement social programs backed by valid evidence of effectiveness, or (ii) sponsor or conduct an evaluation to determine whether a program is effective. The guide provides a brief overview of which studies can produce valid evidence about a program's effectiveness. The final section identifies resources for readers seeking more detailed information or assistance.

I. Well-conducted randomized controlled trials (RCTs), when feasible, are widely regarded as the strongest method for evaluating a program's effectiveness, per evidence standards articulated by the Institute of Education Sciences (IES) and National Science Foundation (NSF),[1] National Academy of Sciences,[2] Congressional Budget Office,[3] U.S. Preventive Services Task Force,[4] Food and Drug Administration,[5] and other respected scientific bodies.

A. Definition of RCT: A study that measures a program's effect by randomly assigning a sample of individuals or other units (such as schools or counties) to a "program group" that receives the program, or to a "control group" that does not.

For example, suppose that a government agency wants to determine whether a job training program for offenders being released from prison is effective in increasing their employment and earnings, and reducing recidivism. The agency might sponsor an RCT which randomly assigns such ex-offenders to either a program group, which receives the program, or to a control group, which receives the usual (pre-existing) services for ex-offenders. The study would then measure outcomes, such as employment, earnings, and re-arrests, for both groups over a period of time. The difference in outcomes between the two groups would represent the effect of the new program compared to usual services.

B. The unique value of random assignment: It enables one to determine whether the program itself, as opposed to other factors, causes the observed outcomes.

Specifically, the random assignment process, if carried out with a sufficiently large sample, ensures to a high degree of confidence that there are no systematic differences between the program group and control group in either *observable* characteristics (e.g., income, ethnicity) or *unobservable* characteristics (e.g., motivation, psychological resilience, family support). Thus, any difference in outcomes between the two groups can be confidently attributed to the program and not to other factors.

By contrast, studies that compare program participants to a group of nonparticipants selected through methods other than randomization (i.e., "quasi-experiments") always carry an element of uncertainty about whether the two groups are similar in unobservable characteristics such as motivation. This can be a problem, for example, for studies in which program participants volunteer for the program (indicating a degree of motivation to improve), and are being compared to non-participants who did not volunteer (potentially indicating a lower level of motivation). Such studies cannot rule out the possibility that participant motivation, rather than the program itself, accounts for any superior outcomes observed for the program group.

C. For this reason, recent IES/NSF Guidelines recommend: "Generally and when feasible, [studies of program effectiveness] should use designs in which the treatment and comparison groups are randomly assigned" – i.e., RCTs. 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]

II. To have strong confidence in a program's effectiveness, one would generally look for the replication of positive RCT findings across different real-world implementation sites.

A. Specific items to look for:

1. **The program has been demonstrated effective, through well-conducted RCTs, in more than one site of implementation.** Such a demonstration might consist of two or more RCTs conducted in different implementation sites, or alternatively one large multi-site RCT.

2. **The RCT(s) evaluated the program in the real-world community settings and conditions where it would normally be implemented** (e.g., community drug abuse clinics, public schools, job training program sites). This is as opposed to tightly-controlled ("efficacy") conditions, such as specialized sites that researchers set up at a university for purposes of the study, or other settings where the researchers themselves are closely involved in program delivery.

B. Why strong confidence requires such evidence: Less rigorous evidence, while valuable for identifying *promising* programs, too often is reversed in subsequent, more definitive research.

Reviews in different areas of medicine have found that 50-80% of positive results in phase II studies (mostly small efficacy RCTs, or quasi-experiments) are overturned in larger, more definitive replication RCTs (i.e., phase III).[6] Similarly, in education policy, programs such as the Cognitive Tutor, Project CRISS, and LETRS teacher professional development – whose initial research findings were promising (e.g., met IES's What Works Clearinghouse standards) – have unfortunately not been able to reproduce those findings in large replication RCTs sponsored by IES.[7,8,9] In employment and training policy, positive initial findings for the Quantum Opportunity Program, and Center for Employment Training – programs once widely viewed as evidence based – have not been reproduced in replication RCTs sponsored by the Department of Labor.[10,11] A similar pattern occurs across other diverse areas of policy and science where rigorous RCTs are carried out.

III. When an RCT is not feasible, quasi-experiments meeting certain specific conditions may produce comparable results, and thus can be a good second-best alternative.

A. 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 erroneous estimates of the program's effect in the study sample.[12]

B. We have published a brief summarizing which quasi-experimental designs are most likely to avoid such threats and thus produce valid estimates of impact.

The brief summarizes findings from "design-replication" studies, which have been carried out in education, employment/training, welfare, and other policy areas to examine whether and under what circumstances quasi-experimental methods can replicate the results of well-conducted RCTs. Three excellent systematic reviews have been conducted of this design-replication literature – Bloom, Michalopoulous, and Hill (2005)[13]; Glazerman, Levy, and Myers (2003)[14]; and Cook, Shadish, and Wong (2008).[15] Our brief draws on findings from both the reviews and the original studies.

C. What follows is an overview of the key concepts in the brief:

1. **If the program and comparison groups differ markedly in demographics, ability/skills, or behavioral characteristics, the study is unlikely to produce valid results.** Such studies often produce erroneous conclusions regarding both the size and direction of the program's impact. This is true even when statistical methods such as propensity score matching and regression adjustment are used to equate the two groups. In other words, if the two groups differ in key characteristics *before* such statistical methods are applied, applying these methods is unlikely to rescue the study design and generate valid results.

   As Cook, Shadish, and Wong (2008) observe, the above finding "indicts much of current causal [evaluation] practice in the social sciences," where studies often use program and comparison groups that have large differences, and researchers put their effort into causal modeling and statistical analyses "that have unclear links to the real world."

2. **The quasi-experimental designs most likely to produce valid results contain all of the following elements:**

   - **The program and comparison groups are highly similar in observable pre-program characteristics, including:**

     - **Demographics** (e.g., age, sex, ethnicity, education, employment, earnings).

     - **Pre-program measures of the outcome the program seeks to improve.** For example, in an evaluation of a program to prevent recidivism among offenders being released from prison, the offenders in the two groups should be equivalent in their pre-program criminal activity, such as number of arrests, convictions, and severity of offenses.

     - **Geographic location** (e.g., both are from the same area of the same city).

   - **Outcome data are collected in the same way for both groups – e.g., the same survey administered at the same point in time to both groups.**

   - **Program and comparison group members are likely to be similar in motivation – e.g., because the study uses an eligibility "cutoff" to form the two groups.** Cutoff-based studies – also called "regression-discontinuity" studies – are an example of a quasi-experimental design in which the program and comparison groups are likely to have similar motivation. In such studies, the pro*g*ram group is comprised of persons just above the threshold for program eligibility, and the comparison group is comprised of persons just below (e.g., families earning $19,000 per year versus families earning $21,000, in an employment program whose eligibility cutoff is $20,000). Because program participation is not determined by self-selection, and the two groups are very similar in their eligibility score, there is reason to believe they are also similar in motivation.

   - **Statistical methods are used to adjust for any minor pre-program differences between the two groups** – methods such as propensity score matching, regression adjustment, and/or difference in differences.

- **Preferably, the study chooses the program and comparison groups "prospectively" – i.e., *before* the program is administered.**

  If the program and comparison groups are chosen by the researcher *after* the program is administered ("retrospectively"), the researcher has an opportunity to choose among numerous possible program and comparison groups. For example, the researcher might select a group of program participants from community A or community B, from years 2007 or 2008, or from age-group 16-20 or 20-24; and might select a comparison group from community A or B or other communities in the county, state, or nation. Each of these choices would likely yield a somewhat different estimate of the program's effect. Thus, a researcher hoping to demonstrate a program's effectiveness can often try many different combinations of program and comparison groups and, consciously or unconsciously, select those that produce the desired result, even in cases where the true program effect is zero. Furthermore, it is generally not possible for the reader of such a study to determine whether the researcher used this approach.

  For this and other reasons, retrospective quasi-experimental studies are regarded by social policy evaluation experts, such as Cook, Shadish, and Wong (2008), and scientific authorities, such as the National Cancer Institute and Food and Drug Administration,[16] as providing less confidence than prospective quasi-experiments and RCTs (where the composition of the program and control or comparison groups are fixed in advance). Their susceptibility to investigator bias may make them particularly unreliable when the researcher has a financial stake in the outcome.

IV. Resources we have developed for readers seeking more detailed information or assistance:

- Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence, 2010 (.pdf, 6 pages + cover)

- Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?, 2014 (.pdf, 3 pages + appendix)

- Practical Evaluation Strategies for Building a Body of Proven-Effective Social Programs: Suggestions for Research and Program Funders, 2013 (.pdf, 9 pages)

- Open online workshop: How to Read Research Findings to Distinguish Evidence-Based Programs from Everything Else

- Help Desk: The Coalition offers brief expert advice in evidence-based reform, without charge.

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] 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.

[3] *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.

[4] 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.

[5] 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.

[6] 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. Michael L. Maitland, Christine Hudoba, Kelly L. Snider, and Mark J. Ratain, "Analysis of the Yield of Phase II Combination Therapy Trials in Medical Oncology," *Clinical Cancer Research*, vol. 16, no. 21, November 2010, pp. 5296-5302. Jens Minnerup, Heike Wersching, Matthias Schilling, and Wolf Rüdiger Schäbitz, "Analysis of early phase and subsequent phase III stroke studies of neuroprotectants: outcomes and predictors for success," *Experimental & Translational Stroke Medicine*, vol. 6, no. 2, 2014.

[7] Campuzano, L., Dynarski, M., Agodini, R., and Rall, K. (2009). *Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts—Executive Summary* (NCEE 2009-4042). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

[8] James-Burdumy, S., Deke, J., Lugo-Gil, J., Carey, N., Hershey, A., Gersten, R., Newman-Gonchar, R., Dimino, J., and Haymond, K. (2010). *Effectiveness of Selected Supplemental Reading Comprehension Interventions: Findings from Two Student Cohorts* (NCEE 2010-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

[9] Garet, Michael S. et. al., *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

[10] Schirm, Allen, Elizabeth Stuart, and Allison McKie, *The Quantum Opportunity Program Demonstration: Final Impacts*. Submitted by Mathematica Policy Research to the U.S. Department of Labor's Employment and Training Administration, July 2006.

[11] Miller, Cynthia, Johannes M. Bos, Kristin E. Porter, Fannie M. Tseng, and Yasuyo Abe, *The Challenge of Repeating Success in a Changing World: Final Report on the Center for Employment Training Replication Sites*. Submitted by MDRC to the U.S. Department of Labor's Employment and Training Administration, September 2005.

[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] Howard S. Bloom, Charles Michalopoulos, and Carolyn J. Hill, "Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects," in *Learning More From Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, 2005, pp. 173-235.

[14] Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental Replications of Social Experiments: A Systematic Review," Mathematica Policy Research discussion paper, no. 8813-300, September 2002. The portion of this review addressing labor market interventions is published in "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science,* vol. 589, September 2003, pp. 63-93.

[15] Thomas D. Cook, William R. Shadish, and Vivian C. Wong, "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons," *Journal of Policy Analysis and Management,* vol. 27, no. 4, 2008, pp. 724-50.

[16] Gary Taubes and Charles C. Mann, "Epidemiology Faces Its Limits," *Science*, vol. 269, issue 5221, July 14, 1995, pp. 164-169. Among other things, this journal article contains a clear description of the issue by Robert Temple, Director of the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration: "The great thing about a [prospective control or comparison-group study] is that, within limits, you don't have to believe anybody or trust anybody. The planning for [the study] is prospective; they've written the protocol before they've done the study, and any deviation that you introduce later is completely visible." By contrast, in a retrospective study, "you always wonder how many ways they cut the data. It's very hard to be reassured, because there are no rules for doing it" (p. 169).