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

Coalition for Evidence-Based Policy
A NONPROFIT, NONPARTISAN ORGANIZATION

Originally published
January 2014

Updated by the Laura and John Arnold Foundation's Evidence-Based Policy team
December 2018

This publication was produced by the Coalition for Evidence-Based Policy, with funding support from the William T. Grant Foundation and U.S. Department of Labor.

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

Brief Overview:

Which Comparison-Group ("Quasi-Experimental") Studies Are Most Likely to Produce Valid Estimates of a Program's Impact?

I. A number of careful investigations have been carried out to address this question.

Specifically, a number of careful "design-replication" studies have been carried out in education, employment/training, welfare, and other policy areas to examine whether and under what circumstances non-experimental comparison-group methods can replicate the results of well-conducted randomized controlled trials.

These studies test comparison-group methods against randomized methods as follows. For a particular program being evaluated, they first compare program participants' outcomes to those of a randomly-assigned control group, in order to estimate the program's impact in a large, well-implemented randomized design – widely recognized as the most reliable, unbiased method of assessing program impact. The studies then compare *the same program participants* with a comparison group selected through methods other than randomization, in order to estimate the program's impact in a comparison-group design. The studies can thereby determine whether the comparison-group estimates replicate the benchmark estimates from the randomized design.

These design-replication studies have been carried out by a number of leading researchers over the past 20 years, and have tested a diverse range of non-experimental comparison-group designs.

II. Three excellent systematic reviews have been conducted of this design-replication literature; they reached largely similar conclusions, summarized as follows.

The reviews are Bloom, Michalopoulous, and Hill (2005)[1]; Glazerman, Levy, and Myers (2003)[2]; and Cook, Shadish, and Wong (2008)[3]; their main findings include:

A. If the study compares program participants to non-participants who differ markedly in demographics, ability/skills, or other characteristics, it 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 the study tries to equate the two groups using statistical methods such as regression (to adjust for pre-program differences between the two groups) or matching (to identify and compare subsamples of participants and non-participants who have similar characteristics). In other words, if the participants and non-participants 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."

B. The comparison-group designs most likely to produce valid results contain all of the following elements:

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

   - **Demographics** (e.g., age, sex, ethnicity, educational attainment, employment status, 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).

2. **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;**

3. **Program and comparison group members are likely to be similar in motivation.** One type of comparison-group design in which the two groups are likely to have similar motivation is a cutoff-based study, also known as a "regression-discontinuity" study. In such studies, the program 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. There are other types of comparison-group designs in which the two groups are likely to have similar motivation.[4]

   However, many comparison-group designs use a program group comprised of persons who volunteer for the program, and a comparison group comprised of non-volunteers. In such studies, the two groups are unlikely to be similar in motivation, as the act of volunteering signals a degree of motivation to improve (which could then lead to superior outcomes for the program group even if the program is ineffective).

4. **Statistical methods, such as matching or regression adjustment, are used to adjust for any minor pre-program differences between the two groups**. Although such methods are highly useful in improving a study's impact estimates, no one method performed consistently better than the others across the various design-replication studies.

C. The three reviews reach varying conclusions about whether comparison-group studies meeting the preferred conditions above can consistently produce valid results, replicating the results of large, well-conducted randomized controlled trials. Consistent with Cook, Shadish, and Wong (2008), we believe additional design-replication studies, testing the most promising comparison-group designs against benchmark randomized controlled trials, are needed to convincingly answer that question.[5] What is clear, however, is that meeting the preferred conditions above greatly increases the study's likelihood of producing valid results.

D. Subsequent design-replication evidence has strengthened the case for cutoff-based comparison-group designs as a valid alternative when a randomized trial is not feasible. Such designs are described above (see B.3). Shadish et. al. (2011)[6], summarizing the most recent design-replication evidence, conclude as follows: "First, the [design-replication evidence] generally supports the hypothesis that the [cutoff-based design] produces similar causal estimates to the randomized experiment across the majority of comparisons attempted. Second, whether tested by statistical or practical significance, a nontrivial percentage of the[se] comparisons did not yield the same results." In other words, the cut-off based designs produced impact estimates that were similar to those of the benchmark randomized controlled trials in most, but not all, cases. Chaplin et. al. (2018)[7] reach similar conclusions.

Shadish et. al. emphasize that this somewhat hopeful conclusion applies to cutoff-based designs that limit their sample to persons just above and just below the cutoff for program eligibility (who are most likely to be equivalent in motivation and other characteristics), rather than including persons well above or below that cutoff. The resulting impact estimates thus apply to sample members near the eligibility cutoff, and may not generalize to those further away.

III. Other factors to consider in assessing whether a comparison-group study will produce valid impact estimates:

A. 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 comparison-group 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,[8] as providing less confidence than prospective comparison-group studies and randomized controlled trials (where the composition of the program and control/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.

B. The study follows the same practices that a well-conducted randomized controlled trial follows in order to produce valid results (other than the actual random assignment).

For example, the study should have an adequate sample size, use valid outcome measures, prevent "cross-overs" to or "contamination of" the comparison group, have low sample attrition, use an "intention-to-treat" analysis, and so on.

Appendix:

Sample Form Used to Review a Comparison-Group Study of an Employment and Training Program

Main question to address in your review: Did the study produce scientifically-valid estimates of program impact?

Specific items to be rated by reviewer:

1. **Please assess whether the study produced valid estimates of the program's impact, using the "Brief Overview" document (attached) as a reference.**

   Specifically, please rate the study on a scale of 1 to 5 (5 being strongest, 1 being weakest) on the following categories in the Brief Overview:

	Study Ratings:
The program and comparison groups were highly similar in key pre-program characteristics *before* statistical methods were used to equate the two groups. Please give one composite rating based on items in section 2 of the Brief Overview, including: ▪ Members were selected from the same local labor market. ▪ The two groups were similar in pre-program employment rates and earnings, and demographic characteristics. ▪ Outcome data were collected in the same way, and at the same time, for both groups. ▪ Members of the two groups were likely to be similar in motivation.	
Appropriate statistical methods were used to adjust for any pre-program differences (hopefully minor) between the two groups. Please give one composite rating.	
Preferably, the study chose the program and comparison groups prospectively – i.e., before the program was administered. Please give one composite rating.	
The study had a valid design/implementation in other areas, such as the following. Please give one composite rating. ▪ Adequate sample size ▪ Minimal cross-over, or contamination, between the two groups ▪ Low sample attrition and/or differential attrition ▪ Sample members kept in original group assignment (program or comparison), consistent with intent-to-treat ▪ Valid outcome measures that are of policy or practical importance ▪ Study reports size of effects, and conducts appropriate tests for statistical significance ▪ Study reports effects on all outcomes measured	

   Comment briefly on the reasons behind your ratings.

2. **Based on your ratings and comments above, do you believe this study produced scientifically-valid estimates of program impact?** [Yes / No] Please comment briefly.

References

[1] 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.

[2] 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.

[3] 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.

[4] An illustrative example is a study of a new teaching method in a college course, in which (i) sections of the course that meet on Monday and Wednesday employ the new teaching method while sections that meet in Tuesday and Thursday employ the old teaching method; (ii) students were unaware which sections would employ which method when they enrolled in the course; and (iii) the study estimates the impact of the new teaching method by comparing end-of-semester exam scores in the Monday-Wednesday sections (i.e., program group) to that in the Tuesday-Thursday sections (i.e., comparison group). Because students did not choose their sections knowing whether the new or old teaching method would be used, there is no reason to believe that students in the program group are any more or less motivated or willing to try new methods than students in the comparison group. A study with a design similar to this is Scott E. Lewis and Jennifer E. Lewis, "Departing from Lectures: An Evaluation of a Peer-Led Guided Inquiry Alternative," *Journal of Chemical Education*, vol. 82, no. 1, January 2005, pp. 135-139.

[5] We also strongly support the recommendation of Cook, Shadish, and Wong (2008) that, in design replication studies, the researchers obtaining the comparison-group estimates be kept unaware ("blinded") as to the benchmark randomized estimates they are seeking to replicate. This would rule out the possibility that they consciously or unconsciously choose parameters for the comparison-group design so as to achieve a successful replication (which would cast doubt on the design's ability to produce valid results in real-world application, where benchmark randomized estimates are unavailable).

[6] William R. Shadish, Rodolfo Galindo, Vivian C. Wong, Peter M. Steiner, and Thomas D. Cook, "A Randomized Experiment Comparing Random and Cutoff-Based Assignment," *Psychological Methods*, vol. 16, no. 2, 2011, pp. 179–191.

[7] Duncan D. Chaplin, Thomas D. Cook, Jelena Zurovac, Jared S. Coopersmith, Mariel M. Finucane, Lauren N. Vollmer, and Rebecca E. Morris, "The Internal and External Validity of the Regression Discontinuity Design: A Meta-analysis of 15 Within-Study Comparisons," *Journal of Policy Analysis and Management*, vol. 37, no. 2, spring 2018, pp. 403-429.

[8] 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).