

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



A NONPROFIT, NONPARTISAN ORGANIZATION

November 2009

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.

This publication is in the public domain. Authorization to reproduce it in whole or in part for educational purposes is granted.

We welcome comments and suggestions on this document (jbaron@coalition4evidence.org).

Brief Overview:

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

1. 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 large, well-implemented randomized controlled trials. These studies first compare participants in a particular program with a *control* group, selected through randomization, in order to estimate the program’s impact in a randomized study design. 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 difference between the two estimates represents the bias produced by the comparison-group 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 using various statistical methods to adjust for differences between the program and comparison groups (e.g., propensity score matching, difference-in-differences, ordinary least squares).

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

The reviews are Bloom, Michalopoulos, and Hill (2005)¹; Glazerman, Levy, and Myers²; and Cook, Shadish, and Wong (2008)³; their main findings include:

- **If the program and comparison groups differ markedly in key pre-program characteristics (e.g., demographics, employment), 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.
- **Importantly, 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.

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 Often Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons,” *Journal of Policy Analysis and Management*, vol. 27, no. 4, 2008, pp. 724-50.

- **The comparison-group designs most likely to produce valid results contain all of the following elements:**
 - (i) **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).
 - (ii) **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;**
 - (iii) **Program group and comparison group members are likely to be similar in motivation – e.g., both volunteer for a job training program, and are therefore similarly motivated to improve their employment status, but those who volunteer after the program reaches full capacity do not receive training and form the comparison group.**
 - **Among comparison-group studies, “regression-discontinuity” designs are particularly well-suited for constructing program and comparison groups with similar motivation.** Such designs compare a program group comprised of individuals just above the threshold for program eligibility, with a comparison group of individuals 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.
 - (iv) **Statistical methods are used to adjust for differences between the two groups –** methods such as propensity score matching, regression adjustment, and/or difference in differences. 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.
- **Even when all conditions above are met, it’s still unclear whether comparison-group studies can consistently produce valid results,** replicating the results of large, well-implemented randomized controlled trials. The three systematic reviews, as well as the underlying design replication studies, reach somewhat different conclusions on this issue. What is clear, however, is that meeting the five conditions cited above greatly increases the study’s likelihood of producing valid results.

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

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

This is because if the program and comparison groups are chosen by the researcher *after* the program is administered (“retrospectively”), the researcher may consciously or unconsciously

select the two groups so as to generate his or her desired results. Indeed, in many cases there may be dozens of possible program and comparison groups that the researcher can choose from, and if the researcher looks at enough of these, he or she will likely be able to find one such combination that will produce the desired result just by chance.

Prospective comparison-group studies are, like randomized controlled trials, much less susceptible to this problem. In the words of the director of drug evaluation for the Food and Drug Administration, “The great thing about a [randomized controlled trial or prospective 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.”⁴

- **The study follows the same practices that a well-implemented 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.

⁴ Robert J. Temple, Director of the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration, quoted in Gary Taubes and Charles C. Mann, “Epidemiology Faces Its Limits,” *Science*, vol. 269, issue 5221, July 14, 1995, p. 169.

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:
<p>The program and comparison groups were 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:</p> <ul style="list-style-type: none"> ▪ Members were selected from the same local labor market. ▪ The two groups were largely equivalent 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. 	
<p>Appropriate statistical methods were used to adjust for any pre-program differences (hopefully minor) between the two groups. Please give one composite rating.</p>	
<p>Preferably, the study chose the program and comparison groups prospectively – i.e., before the program was administered. Please give one composite rating.</p>	
<p>The study had a valid design/implementation in other areas, such as the following. Please give one composite rating.</p> <ul style="list-style-type: none"> ▪ 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.