Research design and evaluation design: An Introduction to impact inquiry

Frank G. Cookingham, October 2017

The purpose of this document is to introduce the concept of research design and evaluation design to people with little or no research training or experience. There are many types of evaluation designs, but two types are key to planning impact evaluation: experimental designs and quasi-experimental designs. The characteristics of each type are described.

Note on terminology… I make a distinction between 'research design' and 'evaluation design'. In the literature the terms are often interchangeable. But I use 'research design' to refer to the overall m&e plan that is required in a grant application for a multi-year program, and 'evaluation design' for a specific event such as an end-of-program evaluation.

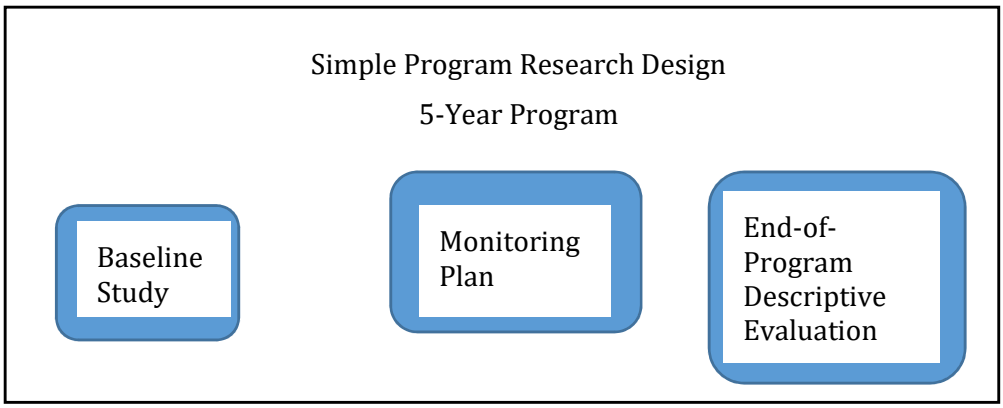
Research Design for Social Program Implementation

In general a RESEARCH DESIGN describes the basic activities that the evaluator will do to collect data and analyze data as a program is planned and implemented. This includes answering research questions, testing hypotheses, exploring how variables react to various conditions, determining program impact, comparing actual program implementation with planned implementation, etc.

Wikipedia describes research design as "the framework that has been created to find answers to research questions." In program evaluation work such a framework includes all of the needs assessment done prior to program planning, baseline studies, implementation monitoring, mid-program evaluation, end-of program evaluation, and post-program sustainability studies. A research design is a plan for collecting and interpreting evaluation data from program "cradle to grave" and beyond.

Evaluation Design

EVALUATION DESIGN is a detailed description of a plan for a specific evaluation exercise at a specific point in the research design that covers the entire life of the program. In other words, the overall research design is a group of evaluation designs. The figure shows a program research design that includes three evaluation designs: baseline study, monitoring plan, and end-of-program evaluation.



The primary function of an evaluation design is to anticipate possible alternative explanations for results, and describe data collection and analysis methods to render implausible as many of them as feasible, leaving a stronger case that results can be attributed to program implementation.

Evaluation of Program Impact

This type of program evaluation aims to provide a plausible estimate of net effects of implementing a program. There are four critical measurements to be completed to obtain the estimate of net effects.

- Gross outcomes for program participants
- Gross outcomes for program nonparticipants who are like program participants in every other relevant aspect
- Degree to which confounding factors have influenced the measurement results of the two groups
- Degree to which evaluation design error and measurement error have influenced measurement results of the two groups

The experimental evaluation design minimizes design error, but in many situations it cannot be used. Quasi-experimental designs reduce various design errors.

Experimental Evaluation Design. Randomized Evaluation Design

An EXPERIMENTAL evaluation design involves comparing outcomes for program participants with outcomes for non-program participants, where people have been assigned to the program group and the non-program control group at random. For community development programs this can be achieved in theory by assigning communities in an area to two groups at random; program interventions are implemented in only one of the groups.

An experimental evaluation design has these basic activities:

- Formulating the evaluation questions and propositions to be studied. This includes reviewing relevant literature and conversations with stakeholders: What do you want to know; how will you use that knowledge?
- Identifying factors that may obscure or confound the effects of the evaluation variables, and planning data collection and analysis that could rule out those factors as causes of the measurements and observations obtained.

- Selecting a sample of study participants from the target population. Ideally the sample of participants is selected randomly, not haphazardly. Often this is not feasible, so the sample is selected by convenient means. This limitation and probable consequences should be discussed in the evaluation report.
- Using a random-assignment procedure to assign people in the sample to two or more groups, at least one of which serves as a "control" group. People in the control group do not experience the conditions experienced by the people in the other groups other than being observed or measured in the same way.
- Administering the experimental interventions specified in the evaluation design, and doing the observations and measurements as prescribed. That is, implementing the planned program, documenting indicator results in the monitoring system, and collecting data as specified in the end-of-program evaluation design. In the end-of-program evaluation the same information is collected from all groups in the same way.
- Applying statistical analysis techniques to the obtained measurements and observations.
- Interpreting the statistical analysis results, and applying other analysis techniques.
- Following the agreed process for reporting the findings of the study.

The formula for estimating net program effects:

> Net Effect = [Gross Outcome for Program Group] – [Gross Outcome for Control Group] ± [Design effects and random error]

Quasi-experimental Evaluation Design

A QUASI-EXPERIMENTAL evaluation design is used when random assignment to groups, especially a control group, cannot be done for ethical or practical reasons. But interpretation of results requires comparison of those results with some credible estimate of results that a similar population would have without participating in the same program. The quasi-experimental design approximates an experimental design in that there is a program group and a non-program group (control group). The absence of random assignment, however, allows a number of extraneous factors to influence outcome measures, which can make it difficult to decide what the program effect is without influence by those factors. Hence the term "quasi-experimental", which means having some but not all features of "experimental."

The evaluator, over the life of the program, and in some cases for periods of time before and after the program, collects and analyzes data to show the strength of common threats to the validity of the program effects (results or impact), and to estimate results for a counterfactual group. This is accomplished by using repeated measures on program participants before and after they experienced program interventions, or comparing participant results with the same measures for a group of people matched with participants on relevant variables, or using statistical controls, or using norms.

This is the formula for estimating program net effects for a quasi-experimental evaluation design.

> Net Effect = [Gross Outcome for Program Group] – [Gross Outcome for Constructed Control Group] ± [Uncontrolled difference between the two groups] ± [Design effects and random error]

The various quasi-experimental designs are different ways of compensating for uncontrolled differences between the program group and a constructed control group.

An evaluator who has little or no experience with quasi-experimental designs should consult with an experienced evaluator or consultant.

Definitions of concepts that undergird experimental and quasi-experimental designs

COUNTERFACTUAL

A conditional statement the first clause of which expresses something contrary to fact, as "If I had known." Retrieve from http://www.dictionary.com/browse/counterfactual. The fact is, the true condition is, that you did not know. So the clause is counter to fact, not the true condition.

In program evaluation the "counterfactual" condition is a group of people who were equivalent in all ways relevant to analyzing program effects prior to program intervention, but did not participate in the program. It is possible that things outside the program may cause positive changes in the outcome indicator; the counterfactual group or condition provides an estimate for such change.

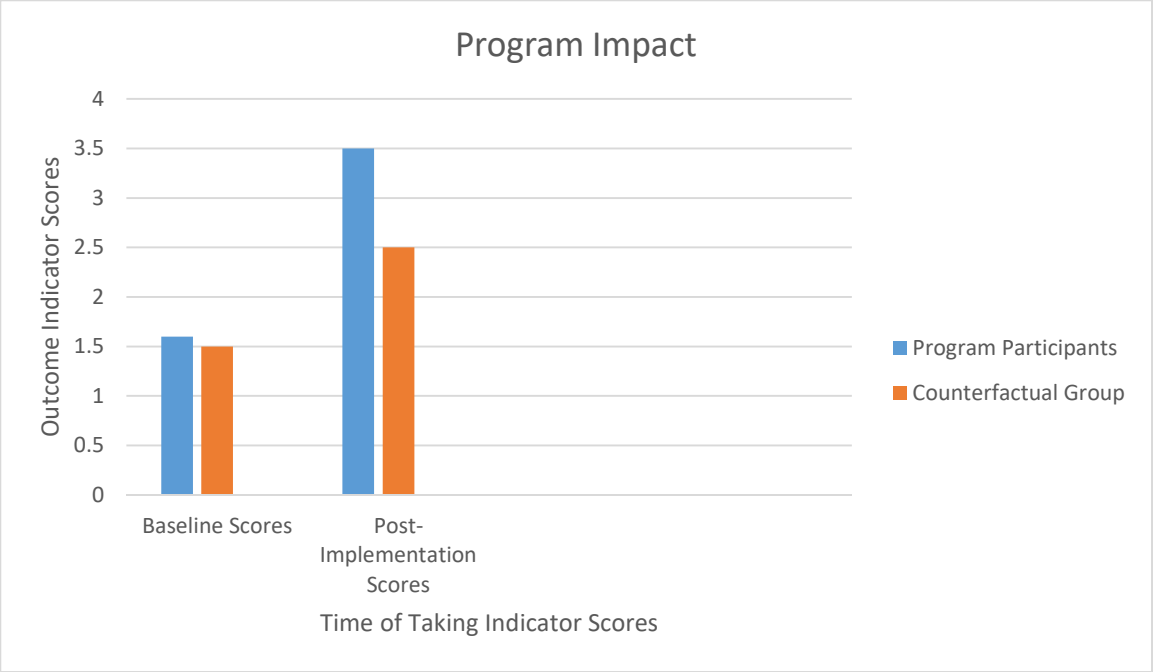
IMPACT

"Impact" is often used as a synonym for "results". But were the results caused by the program, or would similar results have occurred if there had been no program? This is an important issue when a program is costly, or when resources allocated to program design and implementation could be allocated to other uses known to cause desirable change that cannot be achieved as effectively or efficiently by other means.

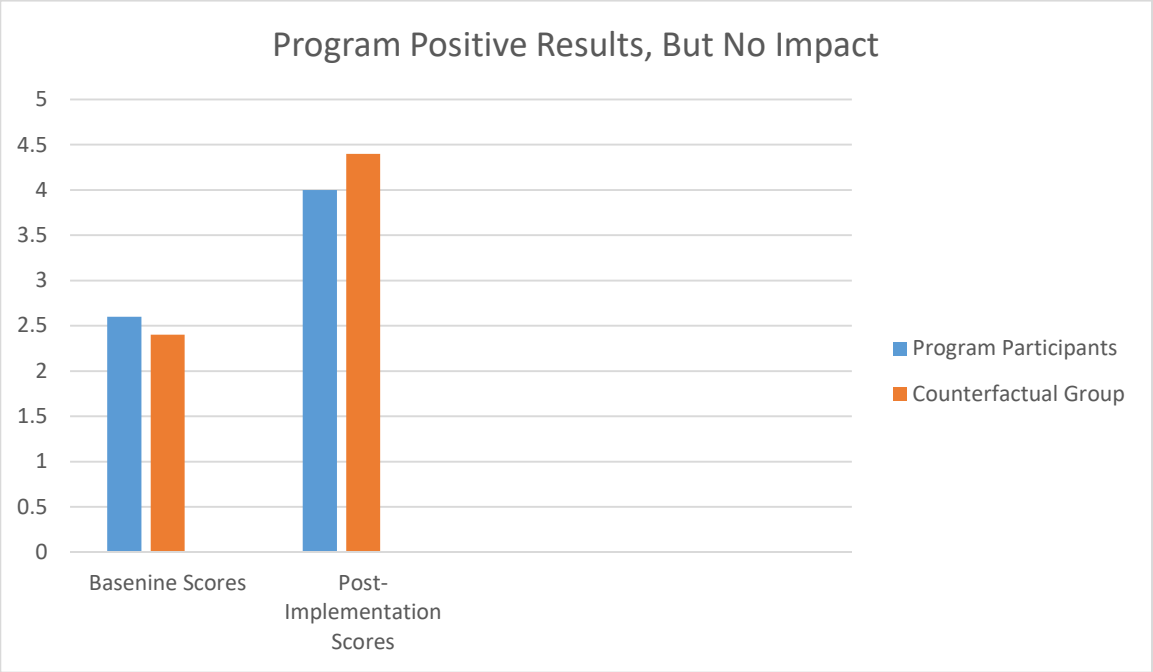
Did the program have the desired impact? In program evaluation this is a short form of asking: Did the program have the desired results, and were those results caused by the program? To answer this question with a "yes" you need to demonstrate two things: (1) positive change in the program group, and (2) less positive change in the control/counterfactual group.

- Positive change in the program group… outcome measures post-intervention for program participants need to be better than outcome baseline measures, assuming that trustworthy methods for data collection and data analysis have been used. In short, there were positive results for program participants. Measured change is necessary to demonstrate impact, but something else is also required.
- Less positive change in the counterfactual group… outcome measures post-intervention for program participants need to be better than estimated outcome measures for a comparison group that was equivalent in relevant ways to the group of program participants, but did not participate in the program. This comparison group is called the counterfactual. This is the evidence that is necessary to conclude that the program caused the results, again assuming that methods for trustworthy data collection were used.

Two charts illustrate this critical point; study them carefully. Describe the difference between the two charts in your own words; is your description equivalent to the two statements above?



Post-implementation scores for both groups showed positive results, but the program participants change results are greater than the counterfactual group results. Both requirements for demonstrating impact are met.



In this chart the change results for the counterfactual group are greater than the results for the program participants. Something other than the program is causing changes in the outcome indicator results.

RANDOMNESS, RANDOM ASSIGNMENT

The concept of 'random assignment' is fundamental to understanding research design. But the adjective 'random' is used in different ways. Here are five definitions of the term; Retrieve from http://www.dictionary.com/browse/at--random

> 1. Proceeding, made, or occurring without definite aim, reason, or pattern: the random selection of numbers.
>
> 2. Statistics… of or characterizing a process of selection in which each item of a set has an equal probability of being chosen.
>
> 3. Building Trades…
>
> (Of building materials) lacking uniformity of dimensions: random shingles.
>
> (Of ashlar) laid without continuous courses.
>
> Constructed or applied without regularity: random bond.
>
> 4. Slang… unknown, unidentified, or suspiciously out of place: A couple of random guys showed up at the party.
>
> 5. Odd or unpredictable, often in an amusing way: my totally random life.

From another dictionary: https://en.oxforddictionaries.com/definition/random

> 1. Made, done, or happening without method or conscious decision: 'apparently random violence'
>
> 2. Statistics … Governed by or involving equal chances for each item: 'a random sample of 100 households'

The common notion of 'random' as without method or conscious decision is counter to the statistical or research notion of 'random'. To do random assignment in a research design requires a conscious decision to select an appropriate method for reasonably ensuring that each person in the sample has an equal chance of being assigned to any of the groups in the design. The most common method is to use a random number table or a random number generator. Failure to use such a method in a disciplined manner poses a threat to the validity of statistical analysis results.

References

Thomas D. Cook and Donald T. Campbell, *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Chicago, Rand McNally, 1979.

> There are more recent treatments of the topic, but this is a classic text that is still a valuable resource. See chapter 8 for a discussion of issues around randomized research design.

Design, Monitoring and Evaluation for Peacebuilding, Module 6: Descriptive, Normative, and Impact Evaluation Designs. Retrieve from http://dmeforpeace.org/sites/default/files/M06_PP.pdf.

> Fifty slides that provide basic information. They could be an outline for a multi-day training workshop.

P. H. Rossi, H. E. Freeman & M. W. Lipsey, (1999), *Evaluation: A systematic approach* (sixth edition), Sage, Thousand Oaks, California.

> Chapters 7-10 provide an excellent description of program impact evaluation.