

Study Designs for Program Evaluation



Contents:

Introduction

What kind of evaluation design will meet my needs?

- What do I need to do an experimental or quasi-experimental study?

Overview of Three Categories of Evaluation Designs

- Exploratory Study Designs
- Descriptive Study Designs
- Experimental and Quasi-Experimental Study Designs

Threats to Validity (or Why your Program Might NOT be Causing the Changes You See)

Project STAR

1-800-548-3656

star@JBSinternational.com

www.nationalservicerresources.org

(Search: project star)

Study Designs for Program Evaluation



Introduction

At different points in your program cycle, you may need to use different types of evaluation designs. You can think of evaluation designs in three main categories:

Exploratory evaluation study designs can help you at the beginning of your program to identify what services to provide and the best approaches to providing those services. It can also help you determine what outcomes will be appropriate for you to measure, given the type of services you offer, and the best way to measure them.

Descriptive study designs can help you show whether your program is operating as planned, provide you with feedback about the services you offer, determine whether your program is producing the types of outputs and outcomes you want, and help clarify program processes, goals and objectives.

Experimental and quasi-experimental study designs can help provide more evidence of a causal or correlational relationship between your services and the outcomes you measure.

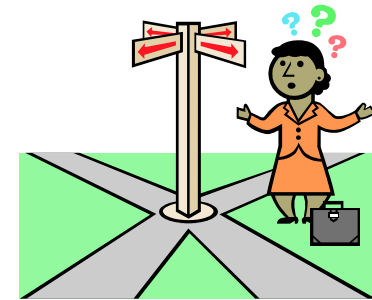
Example: A program conducts a simple needs assessment (*exploratory study*) and learns that the target students may benefit from intensive one-on-one tutoring in math and reading to improve their academic performance and sense of self confidence. A multi-method evaluation (*descriptive study*) including a questionnaire, observation, test, and existing data is used to get information on the intended change, as well as the relations between students' academic performance and self-confidence. A group of students is identified, and with their parents' consent, are randomly assigned to participate in a controlled study (*experimental study*) of whether a good breakfast makes a difference in their performance. The program staff are also interested in knowing whether the added individual counseling and/or support group services help the participating students compared to non-participating students. (*quasi-experimental non-equivalent group studies*).

What type of evaluation design do I need?

The type of evaluation design you choose will depend on the questions you are asking. On the next page is a checklist to help you identify some options.

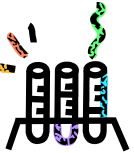
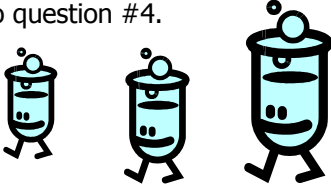
What type of evaluation design will meet my needs?

Use this checklist and the chart on the next page to help you determine what kind of evaluation study will meet your needs. For all of these studies, we recommend that you *contact a professional evaluator* to assist you in planning and implementing the evaluation, and interpreting the results.



<i>Do you need to...</i>	<i>Check if yes</i>	<i>If yes...then you need to do this type of study:</i>
Identify community needs?	<input type="checkbox"/>	Exploratory study: see page 4 for a description.
Identify good service activities for meeting your community's needs?	<input type="checkbox"/>	Exploratory study: see page 4.
Know whether you are meeting output targets?	<input type="checkbox"/>	Descriptive study: see page 5.
Identify areas for improvement and ways to improve your program?	<input type="checkbox"/>	Descriptive study: see page 5
Know whether your program participants increased in knowledge, skills, behavior or attitudes? That is, document changes in beneficiaries (outcomes).	<input type="checkbox"/>	Experimental, quasi-experimental or a descriptive study: see the chart on the next page and page 7.
Evaluate your program and demonstrate program effectiveness? That is, obtain evidence that the program caused the outcomes observed in beneficiaries.	<input type="checkbox"/>	Experimental or quasi-experimental study: see the chart on the next page and page 7.

What do I need to do an experimental or quasi-experimental study?

Do you have ...	Yes/No	
<p>1. The ability to RANDOMLY ASSIGN* participants to either participate in the program or not? For example, do you have a waiting list that you can pull names from randomly? Or, can you ethically not serve some people based ONLY on RANDOM ASSIGNMENT? Or, can you ethically delay service to some RANDOMLY ASSIGNED participants until post-service data can be collected from other participants?</p>	<p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>	<p>If yes, you may be able to use an experimental design.</p> <p>If no, then go to question #2.</p> 
<p>2. A group of people similar to your participants who will <i>not</i> be receiving services but for whom you can get measurements? For example, can you get measurements for students at another similar school, caregivers who are too far away from your location to get respite services, or people on a waiting list who signed up too late to receive your program services?</p>	<p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>	<p>If yes, you may be able to use a quasi-experimental design— matched or non-equivalent control group design.</p> <p>If no, then go to question #3.</p>
<p>3. The ability to collect multiple measures (or access existing data) on your service participants before and after they receive services? This would involve 3 - 6 measures before and after the service, spaced over the course of 1 - 5 years, depending on the length of the service you provide (shorter services may need shorter measurement periods; longer services may need longer measurement periods). Hint: This works better for environmental programs, where baseline data can be collected over long periods of time, or for programs using standardized measures that are accessible to, but usually not administered by the program.</p>	<p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>	<p>If yes, you may be able to use a quasi-experimental design— a time series design.</p> <p>If no, then go to question #4.</p> 
<p>4. Access to a professional evaluator who can assist you in designing a quasi-experimental study?</p>	<p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> No</p>	<p>If yes, there are multiple quasi-experimental statistically-based approaches that may be useful to your program in demonstrating effectiveness.</p> <p>If no, review Project STAR’s “Selecting an Evaluator” packet or contact Project STAR.</p>

* RANDOM ASSIGNMENT means all participants have an equal chance of being selected for the treatment and control groups.

Overview of Three Categories of Evaluation Designs

Exploratory Study Designs



Why use an exploratory study?

The purpose of an exploratory study is to gain familiarity, increase understanding, and to help to formulate better program services, evaluation questions and approaches. This process may involve some or all of the following methods:

- *Review of the current literature.* What do we already know about this topic? For example, what interventions have been successful in the past, with whom, and under what circumstances?
- *Review of existing data or information.* What information has already been documented about the people you are serving? What are their characteristics, resources and limitations? This may include demographic, educational, cultural or other information that can be obtained through your agency, collaborating agencies, or public resources (e.g. census data).
- *Study of selected examples.* This might involve open-ended surveys, interviews or focus groups of a cross-section of people/programs/groups from the target population.
- *Interviews or surveys with individuals with different viewpoints, or with key informants from your stakeholder group.* This allows the researcher to see the topic/intervention in different lights, as well as generate investment in the program development process.

Example: A program conducts a simple needs assessment (exploratory study) using methods including: a review of the literature on tutoring strategies and outcomes, focus groups with teachers, surveys of parents, and interviews with a cross-section of local youth service providers. The program learns that the target students may benefit from an intensive one-on-one tutoring in math and reading to improve their academic performance and sense of self confidence.

Project STAR

Descriptive Study



Why use a descriptive study?

The purpose of a descriptive study is to provide an in-depth description of a phenomenon or the relationships between two or more phenomena. Here are three common goals of a descriptive study:

- Describe service recipient or program characteristics and how they relate to one another (study of correlation).
- Describe the use of community resources (service utilization).
- Solicit views of a group of people on an issue, as in an opinion survey, satisfaction survey or poll.
- Document program processes, outputs and outcomes.

A descriptive study differs from an exploratory study in that there is more attention to securing a representative sample and the study may involve comparison groups. Data-gathering techniques also tend to be more precise in a descriptive study and there is a clearer and more specific focus on what is being studied. Common methods used in descriptive studies include:

- *Analysis of existing data or information.* This can answer questions like: Are we meeting our output targets? Who participates in program services? “What characteristics do participants have in terms of demographics, attitudes, history, and pre and post test scores on any relevant existing measures?”
- *Survey, interview or focus group data relating to program experiences.* This might involve open-ended surveys, interviews or focus groups of a randomly or systematically selected group of program participants. These data collections can be designed to collect data from people who either fairly represent your service population (e.g. are selected using a randomization process) or who are targeted because they can provide useful information (e.g. only participants who have attended all program sessions are surveyed).
- *Preliminary outcome measures or tests.* Pre and post testing/outcome measurement of participants during the descriptive process will allow you to see if your program is operating as intended.
- *Extended statistical analysis of data collected.* Correlation or other types of statistical analysis of data collected from existing sources, surveys, interviews, or preliminary outcome measures can be conducted (usually by a professional evaluator or statistician) to help answer questions about your program’s participants, processes and outcomes.



Project STAR

Descriptive studies aim to provide data for these major evaluation questions:

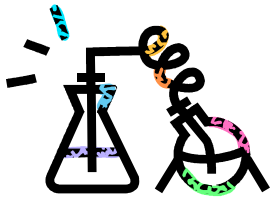
1. Does **X** have a certain characteristic?
 - *Are participants in this program highly motivated to succeed at entry?*
 - *Do first grade students who were read to frequently as preschoolers have an easier time learning to read in this program?*
 - *Are streams with high toxin measurements more likely to be in highly populated areas?*
2. Does **X** occur more frequently than **Y** in a given population?
 - *Does program drop-out occur more frequently than program completion among those who speak Spanish as their primary language?*
 - *Do unexcused program absences occur more frequently than documented illnesses in students who drop out of our program?*
 - *Do invasive species occur more frequently than native species in highly polluted streams?*
3. Is **X** associated with **Y** in some important way?
 - *Do participants with higher rates of attendance have higher outcome test scores?*
 - *Do youth whose parents are more actively involved in our program have higher levels of participation?*
 - *Is recent garbage clean up in the watershed associated with lower toxin levels in the streams we measure?*

Example: A multi-method evaluation (*descriptive study*) including a questionnaire, observation, test, and existing data is used to get information on the intended change as well as the relations between students' academic performance and self-confidence.



Project STAR

Experimental and Quasi-Experimental Study Designs



Experimental and quasi-experimental study designs are often utilized in summative evaluations. Experimental studies try to determine causality or correlation (in so far as this is possible); that is, the change in the dependent variable (your outcome measure) depends on change in independent variable (your service). To increase the likelihood of determining if the change is caused by your service (and not something else), experimental studies compare an *experimental group*—the group that received the intervention/service — to a *control group*, a similar group that did not receive the intervention.

Why use Experimental or Quasi-experimental Study Designs?

Experimental and Quasi-Experimental approaches, when carried out well, can help you:

- Show that your services (not something else that happens to participants at the same time, or some change in the participant’s environment etc.) are contributing to the outcome you are measuring.
- Show that your services would likely cause the same types of outcomes for other potential participants who were not part of your study.

Experimental studies aim to provide data for these kinds of evaluation questions:

- *Contributory*: X increases the likelihood of Y, but X is only one of a number of factors.
Studying increases the likelihood of getting good grades.
- *Contingent*: A condition may have a causative influence on whether X increases the likelihood of Y, under certain circumstances or in certain contingencies.
Job placement services will decrease the number of people who are on the welfare roll, when the economy is doing well.
- *Alternative*: Either X or Z increases the likelihood of Y.
Watershed clean up or low population density will lead to lower toxin levels in streams.

Experimental and quasi-experimental study designs both involve comparing two groups to see if the desired outcomes are more likely to occur in the group that received the intervention. In the experimental study design, however, subjects are *randomly chosen or assigned* to a group.



Though experimental study designs are considered the “gold standard,” they may not be ideal or feasible for both ethical reasons (random assignment means one group does not receive the service/intervention) and practical reasons (a high level of investment is required, including time, expertise, and often expense).

Experimental studies are considered the gold standard for study designs because they do the best job of limiting or eliminating threats to your study’s internal validity. (See *Threats to Validity* on page 10.)

Project STAR

Quasi-experimental studies are often more realistic in service delivery settings. For example, a group of children who receive tutoring can be compared to the other children in their class who did not receive tutoring to see if grades are generally better in the tutored group. However, because randomization is not used, one validity threat to quasi-experimental studies is “selection bias.” It may be that there is something different about the people who choose to participate in the program (e.g. they are more motivated) that makes them more likely to succeed. How can we be sure that the change we see (e.g. improved grades) was caused by the service and not this personal characteristic?

Examples of the Experimental and Quasi-Experimental Designs

In the following examples:

- **X** refers to the administration of the independent variable (the intervention/service).
- **O** refers to the observation of measurement of the dependent variable (the expected output, outcome/change).
- **R** indicates randomization (i.e. random sampling or random assignment). Randomization increases the likelihood that the groups will have the same variance in important characteristics that may affect the outcome. Experimental designs usually involve the use of randomization. (Note: Randomization is different from *random selection*, in which all members of your service population have an equal chance to participate in the experiment, either as treatment or control participants. Random selection helps show that your study results represent the whole population of interest [i.e. everyone in your service population], but is not required to make a study a true experiment.)

In these examples, let’s say **X** is a tutoring service and **O** is a test score that students receive.



A. Experimental Designs (involves random assignment)

1. Pretest-posttest control-group design

R	O₁	X	O₂
R	O₁		O₂

In this design, students are randomly assigned to a group that receives tutoring or to a group that does not receive tutoring. Both groups are tested twice, at the same times; for the experimental group, this is before and after receiving tutoring. Pretest scores for both groups should have a similar distribution so that we know the two groups started in the same place. If the tutoring was effective, the distribution of the post test scores from the group that received tutoring should be higher than the post test scores from the control group.

Project STAR

2. Posttest-only control-group design

R X O
R O



In this design, students are randomly assigned to a group that receives tutoring or to a group that does not receive tutoring. Both groups are tested once, at the same time; for the group that received tutoring, this is after tutoring has occurred. If the tutoring was effective, the distribution of the post test scores from the group that received tutoring should be higher than the post test scores from the control group.

B. Quasi-experimental designs (no random assignment, multiple assessments of change)

1. Non-equivalent comparison group pretest-posttest

O₁ X O₂
O₁ O₂



In this design, two groups of students, one of which receives tutoring, is tested at two points in time. For the group that receives tutoring, tests occur before and after tutoring. Pretest scores for both groups will tell us whether the two groups started in the same place. If the tutoring was effective, the distribution of the post test scores from the group that received tutoring should be higher than the post test scores from the control group. Note that in this case, because students were not randomly assigned to one group or the other, we can not be sure that advances made by the tutored group were due solely to the intervention. It may be that the group that chose to participate in the tutoring program is more motivated than the group that opted out, or their parents are pushing them more to study, etc.

2. Time-series quasi-experimental design / interrupted time-series

O₁ O₂ O₃ O₄ X O₅ O₆ O₇ O₈

In this design, one group of students that receives tutoring is tested at different intervals before receiving tutoring, and again after receiving tutoring. If the tutoring was effective, we would expect to see a significant increase in the scores immediately afterward (O₅); that is, a jump in scores higher than what may have been occurring before tutoring. Ideally, this score will be maintained or even increase slightly at later intervals (O₆, O₇, and O₈).

Threats to Validity

Threats to Validity are factors other than the program services that might be the cause for change. Below is a list of threats to the validity of your results; experimental designs aim to limit these threats.

Internal validity: does your program make a difference for the people you measure?

Factors which jeopardize internal validity	What you need to consider:	For example:
History	Did something happen to effect the outcome, besides your intervention, between the first and second measurement?	How do you show that other things that are happening in the lives of your students are not changing their test scores?
Maturation	What participant improvements would be seen over time, even without any intervention?	How do you show that the increase in reading scores you see isn't simply due to children getting older?
Testing	What are the effects of taking a test on the outcomes of taking a second test?	How do you know that practicing (i.e. pre-testing) on the measures you use isn't helping students improve their scores?
Instrumentation	What changes in the instrument, observers, or scorers might produce changes in outcomes?	How do you know that the change in test scores you see (e.g. from pre to post, or from treatment to control) isn't simply due to a change in your test?
Statistical regression	How do you know if individuals who score very high or very low at one point in time, will likely score closer to the middle at the next measurement? This is called "regression to the mean."	If you select participants into your treatment who have the very worst or the best scores on your pre-test measures, how do you know they are not "regressing to the mean" when their scores improve?
Selection of subjects	How do you know that the way in which you select individuals into treatment or control groups doesn't affect the outcome you get?	How do you know you didn't put only the most motivated students into your tutoring program?
Experimental mortality	How do you know that the individuals who dropped out of the treatment group before the post test were similar to those who dropped out of the control group?	How do you know that the increases you see in your outcome measure aren't due to low scorers dropping out of tutoring?
Selection-maturation interaction	How do you know that the selection of comparison groups and their maturation are not interacting to lead to confounding outcomes, and erroneous interpretation that the treatment caused the effect?	How do you know that the increase you see isn't simply due to differences in normal reading development between children in your tutoring program and those in your control group?

Project STAR

Internal validity threats (cont.)	What you need to consider:	For example:
John Henry effect	John Henry was a worker who outperformed a machine under an experimental setting because he was aware that his performance was compared with that of a machine. How do you know that your treatment group isn't simply putting forth extra effort because they know they are being compared to others?	How do you know the results you see aren't due to your student's desire to "outperform" the comparison group?

External validity: would your service make a difference to other potential participants?

<i>Factors which jeopardize external validity</i>	What you need to consider:	For example:
Reactive or interaction effect of testing	How do you know that your pretest might not be increasing or decreasing a subject's sensitivity or responsiveness to the experimental variable?	How do you know that your program pretest (which is not normally part of the curriculum) isn't helping prepare the students to learn the material?
Interaction effects of selection biases and the experimental variable	How do you know that those selected for your treatment group respond like the rest of your targeted service recipients?	How do you know that the students receiving service will respond the same way as the rest of the students you would like to serve?
Reactive effects of experimental arrangements	How will the outcomes generalize to non-experimental settings, if the effect was attributable to the experimental arrangement of the research?	How do you know that your tutoring program will not be fundamentally changed once it is not being evaluated?
Multiple treatment interference	If multiple treatments are given to the same subjects, what are the effects of prior treatments?	If you are measuring different services using the same group of students, how do you know which of the (separate) services your students receive are causing the change?