Corporation for NATIONAL & COMMUNITY SERVICE

Evaluation Design Tips

Introduction

To ensure you get the strongest possible evidence results from the evaluation, it's important to use the correct evaluation design. This document offers tips on how to select the appropriate evaluation design and describe that design in your SEP. Each section of this document contains information on one of the most common evaluation designs included in the SEP Guidance document.

Experimental Studies

Randomized Between-Groups Design

The first thing to consider when choosing a study design is whether or not you will be able to randomly assign schools, classrooms, and/or students to treatment and control groups. If you can deliberately create randomized treatment and control groups, then there are many important factors to consider.

In creating a between-groups evaluation design formed by randomization, you should ask yourself the following questions:

> ***What will my groups consist of? How will they be randomized?***
>
> Consider who will be in each group. For example, will the groups be composed of program participants and nonparticipants, or of participating and nonparticipating program sites? Be sure to fully explain the process by which randomization will occur. State if randomization will occur before people enter the program, if equal numbers of people will end up in both the treatment and control groups, and if there are any other particular features of the randomization process. How might the evaluation process and evaluator decisions (such as choice of survey items) affect my results?
>
> ***How will that randomization be protected?***
>
> To ensure that the between-groups evaluation design formed by randomization fully accounts for threats to internal validity, you need to consider the ways in which random assignment might be thwarted, whether on purpose or purely by chance. For example, might program staff be resistant to withholding treatment to some, or all, possible participants?
>
> ***Will I stratify my sample to improve the variability in my data or to provide sufficient statistical power to examine particular subgroups of the sample?***
>
> Sometimes it is difficult to ensure that simple random assignment will create

equivalent groups of individuals, programs, or sites. Determine if stratifying the sampling procedure would improve equivalence among the groups or if it would increase your ability to compare specific subgroups within your sample (e.g., comparing men and women or comparing people with certain illnesses).

What will be the control condition?

Determine what people, programs, or sites in the control group will experience. Will control group members be eligible for program participation at a future time; will they be able to take part after a specified period, or will they be directed to some other program or service that is an alternative to your program?

Lastly, even though many standards do not require it, it might think about pre-test characteristics that could be measured to document group equivalence and to potentially account for group nonequivalence in your analysis.

Describing Randomized Between-Groups Design

- Unit of random assignment is clearly identified (and aligned with unit of analysis).

  Describe the unit of randomization: will it be individuals, groups, sites, programs, etc.?

- Procedures to conduct the random assignment, including who implemented the random assignments, how the procedures were implemented, and how the procedures were used to verify that probability of assignment groups, are described and generated by random numbers.

- Blocking, stratification, or matching procedures used—to improve precision in the estimate of the program effect or to balance groups on measured characteristic(s)—are described.

  Present and detail how people (or programs, sites, or groups) will be randomly assigned to either the treatment or control groups. Provide detailed information as to how and when assignment will occur, and who will oversee assignment (i.e., program staff, evaluation staff, or some other party). If blocking, stratification, or matching is used, be sure to include detailed information about those procedures and how they will be implemented successfully.

- The program group and to the extent possible, the control group conditions, are described.

  Drawing on the logic model, describe what constitutes program participation compared to the experiences of the control group.

- Any concerns that proposed strategies or approaches will lead to nonequivalent groups are discussed.

Ideally, random assignment will yield statistically similar treatment and control groups, but implementation of random assignment in the real world does not always happen as planned. Describe procedures that will be used to deal with any non-equivalence of the treatment and control groups in the study.

Between-Groups Design, Formed by Matching or Cut-off Score

On many occasions, random assignment is not feasible. One such circumstance is that programs may have a prior agreement with schools that a program will be implemented in a certain set of schools, and the effectiveness of this program is to be evaluated. Some potential participants may be hesitant to participate when there is a chance that they might be assigned to non-treatment. In such situations, a quasi-experimental design becomes necessary.

In creating a between-groups evaluation design formed by matching or cut-off score, you should ask yourself the following questions:

How can you construct treatment and comparison groups that are as similar as possible?

Remember that, even when pre-test characteristics are controlled for in a QED analysis, this will not account for any unobserved characteristics, or be sufficient when between-group differences are moderate (an effect size of 0.20). So, selection for the comparison group is very important.

Can you minimize the differences between the participant group and the comparison group?

Matching participants in the treatment and comparison groups is always wise, even when the comparison groups have been selected carefully, since you want to minimize the observed differences between the treatment and comparison groups. Propensity scores matching is an effective means of using multivariate models to build comparable treatment and control groups. Bear in mind, when using this technique, that: a) the more pre-test measures available for matching the better; b) multilevel modeling and clustering can be incorporated into propensity scores; and c) a conservative approach is to control for all measures used in matching the calculation of any treatment effect. Treating the comparison group as carefully as the treatment group is very important.

What means can be applied such that rates of participation and attrition are reduced for both the treatment and comparison groups?

As with other research designs that use a comparison or control group, even well-

Evaluation Design Tips, continued

planned research designs can run into unexpected problems during implementation of the design. However, planning for potential problems and developing mitigation strategies, ranging from detailed follow-up plans that ensure data collection after program participation to over-sampling of certain subgroups, can help lessen some of these issues.

To what population can your results be generalized? Would it be possible to diversify your sample such that external validity is improved?

Describe the strengths and weaknesses of the samples included in your study. Determine how representative your program participant and comparison group members are of the target population, and consider using post hoc statistical adjustments to make your results more generalizable.

Describing Between-Groups Design, Formed by Matching

- Unit of matching is clearly identified (and aligned with the unit of analysis).

  Describe the unit of matching: individuals, groups, sites, programs, etc.

- Procedures to carry out the matching to form a comparison group are described.

  Present and detail how people (or programs, sites, or groups) will be matched. Provide detailed information as to how the matching will be done, including a description of any statistical procedures to be used (such as propensity score matching), and who will conduct the matching.

- A precedent in the literature for including the variables used in the matching is included.
- Methods used to form the proposed comparison group are described such that the validity of the matching is explained.

  Draw upon previous research in presenting the validity of the use of the matching procedure. Include evidence that the matching variables are suitable for use with the matching procedure chosen.

- Reasons why the comparison group might differ from the treatment group and threaten internal validity, and the ways in which the proposed methods adjust for those differences, are discussed.

  Describe the strengths and limitations of your matching procedure. Detail procedures that will be used to deal with differences between the program participant and comparison groups in the study. Include in your discussion any preemptive or post hoc procedures that will be used to achieve greater equivalence between the groups.

Evaluation Design Tips, continued

Describing Between-Groups Design, Formed by Cut-off Score (RDD)

- Measure and cut-off score are clearly identified and aligned with the unit of analysis.
- Cut-off score is clearly delineated and justified.
- Methods used to apply the cut-off score are described in detail.

  Describe the measure that will be used for the cut-off score, and present evidence for using the particular score as the cut-off point for the analysis. Include justification for use of an exact or fuzzy cut-off score. Detail the methods that will be used to apply the cut-off score.

Interrupted Time Series

Interrupted time series can be a useful technique when data from only a small number of units (or even a single unit) is available, but when multiple years of pre- and post-test data is available. It is very important, however, that the study have access to data at many time points pre- and post-implementation to sufficiently rule out threats to internal validity. So, it is important to consider whether or not it is possible to collect enough data given the budget and timeline. This may be a particularly pressing issue for pre-test data.

In creating an interrupted time series design, you should ask yourself the following questions:

Is it feasible to measure the outcome of interest several times prior to the implementation of the program?

Describe how data will be (or will have been) collected before, during, and after program participation. Some outcomes, such as calorie intake or heart rate, may be more amenable to this than things such as gang membership or school achievement, which may not show change over short time periods.

Will you use archived data to measure pre-treatment outcomes?

Consider how applicable these data are to your outcome of interest (i.e., a direct equivalent measure or a proxy). Are measures taken from archived data realistically likely to be affected by your program? For instance, is a small anti-bullying effort enrolling a few students likely to impact school rates of suspensions and expulsions?

Do you have comparison groups or non-equivalent dependent variables to eliminate threats to internal validity?

Interrupted time series designs require comparison groups or non-equivalent dependent variables to eliminate threats to internal validity. Therefore, it is

important to consider whether it would be possible to collect data from non-treated individuals, schools, etc. Alternatively, would it be possible to measure similar outcomes pre- and post-test that are not targeted by the intervention? For example, would it be possible to measure alcohol and marijuana use when an intervention targets smoking?

Describing Interrupted Time Series Design

- Number of measurement points before and after the intervention is described.
- Number of measures during each measurement phase is shown to be sufficient to establish a trend and rule out rival explanations.
- Timing of measures pre/post interruption is shown to be appropriate to the intervention.
- Comparison cases are clearly described.

  Describe the number and frequency of measurements of outcomes data. Include information that indicates that the number and frequency of measurements is suitable for the analysis being undertaken. If a comparison group is included to strengthen external validity, be sure to describe how measurements will be taken for that group as well.

Pre-Experimental Designs

Pre-experimental designs can provide a plethora of useful preliminary information for a program in its early stages of development, and this information can be particularly vital as program scale up to wider implementation and more rigorous evaluation. For example, a feasibility study might test new measures, instruments, and data collection techniques, but proposals should be clear as to the impact these tests will have on future evaluations of this and other programs. It is important to ask whether the program to be tested is at a very early stage in its development. Therefore, what are the barriers to a more rigorous evaluation? What will be the steps toward evaluating the effectiveness of the program more properly, and how does this project play a key role?

In creating a pre-experimental design, you should ask yourself the following questions:

Is this program so novel that no precedent for its implementation has been set? Has no program of this type ever been attempted?

Describe the reasons why a pre-experimental study is appropriate for this program. Include evidence of distinctness of the program, or ways in which the program dramatically changes an existing program.

Evaluation Design Tips, continued

How will the present study develop new measures or instruments for data collection? What improvements can be made over existing measures?

Include an assessment of how the current evaluation will help generate future evaluations that are likely to yield causal evidence of program efficacy. Detail the ways in which current data collection, measures, and sampling will improve future efforts to assess the program. Describe whether or not the current study will lead to the development of a data collection system that will facilitate future evaluations (e.g., an online survey collection procedure might be developed).

Describing Pre-Experimental Design

- Barriers to proposing a design with the potential to contribute to strong or moderate evidence categories are described.

  Include the reasons that a research design that would yield moderate or strong evidence is not appropriate given the particular characteristics of the program. Situate your discussion within the context of any previous research done on the program, the stage of program implementation, and any information about program changes due to expansion or transformation.

- Full study design is clearly and comprehensively explained.
- Additional threats to the internal validity of the design are discussed.
- Ways future evaluations can be designed to rule out these threats are described.

  Describe the study design in detail, noting how threats to internal validity are being addressed when possible. Include any ways in which the current evaluation will help future evaluations minimize threats to internal validity.

- Description of the treatment and counterfactual groups are included.
- Where appropriate, assignment of study participants to groups is described.

  If possible, include a comparison group, or other counterfactual situation, in the research design. Describe who will be in the program participant group (and the comparison group, if appropriate).

Implementation Evaluation

When programs are not implemented properly, they are not likely to be as effective as they can be. Therefore, ask yourself the following questions:

Evaluation Design Tips, continued

Is there any way that your program might not be implemented the way it was intended?

It is important to consider what barriers people might face in implementing a program. For example, instructors might change materials, or may not be properly engaging students during lessons. Another way that fidelity is affected is that material is delivered by personnel who are not qualified or who have not been fully trained. For example, sometimes teachers leave a school because of illness, and untrained supply teachers deliver program materials.

How might barriers to full implementation or any other such situations be avoided?

Thinking about potential barriers to implementation prior to either implementation of the program or implementation evaluation takes place can help reduce the likelihood that such barriers will become insurmountable. One important means is to provide direct support to program sites, and to collect data on the quality of implementation to monitor fidelity, as well as reinforce the importance of fidelity among vested parties.

How can data on implementation fidelity be integrated into your study?

Implementation findings can be a vital component of feasibility studies as an outcome (under what circumstances was fidelity particularly poor?), as a moderator in outcome evaluations, or as a control variable in process investigations (or exploratory research questions). For instance, was the program effectiveness higher when fidelity was high?

How might the present study serve to inform future implementations of the program? Will you more fully develop training materials and lesson plans? Will you train support staff that will directly facilitate implementation in future?

Consider the ways in which the current evaluation will not only address current implementation of the program, but also serve as a way to potentially improve future expansion of the program.

Measures of fidelity are a particularly important part of feasibility studies. Examples of questions that might be addressed by feasibility studies are: 1) How might data from a preliminary study be used to identify areas where the program can be improved so that it can be implemented easily? and 2) What portions of the program were not utilized, or changed, and why? Using focus groups, or other forms of qualitative data, are frequently a good way to collect fidelity data.

Evaluation Design Tips, continued

Describing Implementation Evaluation

- Specific plans for measuring fidelity of program implementation (i.e., how well the program was actually implemented) in the program group are presented.

  Describe in detail the plans for measuring successful implementation of the program. Consider the benchmarks that will be used to assess fidelity, including previous implementation studies of the program in different venues or of related, but not identical, programs.

- Plans for measuring the level of program services the program group actually received, including the criteria for assessing whether an adequate amount and quality of the program was delivered to participants, are described.

  Include information on how data will be collected related to implementation and on the measures that will be used to assess fidelity. Consider the amount of program services participants will receive to estimate the "dosage," or level of program involvement, among participants.

- Plans for assessing whether the control or comparison group received program services, including the criteria for assessing the extent to which there was diffusion of the program to the control or comparison group are provided.

  If the research design includes a control or comparison group, include information on what services members of that group received during the same period as program participants. Including this information will improve the understanding of how program components specifically affect program participants versus control or comparison group members.