IPDET

Handbook

Module 6

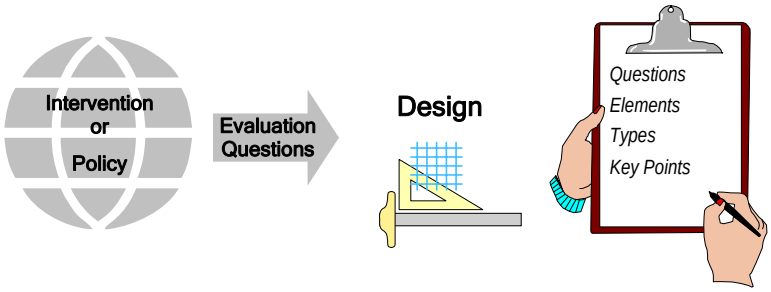
Descriptive, Normative, and Cause-Effect Evaluation Designs

Introduction

After you determine your evaluation questions, you will select evaluation design features or approaches that are most appropriate given each question and situation. This module presents some guidelines, along with the strengths and weaknesses of various approaches, but it is important to keep in mind that every situation is unique. There is no "one and only" way to meet every evaluation situation.

This module has five topics. They are:

- What is Evaluation Design?
- Connecting Questions to Design
- Design Elements
- Types of Designs for Cause-Effect Evaluations
- Key Points about Design.



Learning Objectives

By the end of the module, you should be able to:

- define evaluation design
- describe how evaluation designs connect to evaluation questions
- describe the elements of cause-effect evaluation design, including: before and after measures, comparison groups, random assignment, use of control variables, use of natural variation, and causal tracing strategies
- list and describe the causal tracing strategies
- describe the types of cause-effect evaluation design, including: experimental design, quasi-experimental designs, and non-experimental design
- list and describe the kinds of quasi-experimental design, including: matched, non-equivalent groups, correlational, cross-sectional, interrupted time series, longitudinal, and panel design
- list and describe the types of non-experimental design, including: cross-sectional, time-series, descriptive case studies, before and after, and one-shot.

Key Words

You will find the following key words or phrases in this module. Watch for these and make sure that you understand what they mean and how they are used in the course.

evaluation design
one-shot designs
cross-sectional designs
before and after designs
time-series designs
case studies designs
internal validity
single- or double-blind studies
baseline
comparison groups
treatment group
control group
random assignment
control variables
natural variation
causal tracing strategies
causal list inference
modus operandi (MO)
temporal precedence
constant conjunction
contiguity of influence
strength of association
biological gradient
coherence
analogy
experimental design
quasi-experimental design
correlational design

What is Evaluation Design?

Evaluation design is the total process of specifying a plan for:

- collecting data
- analyzing data
- reporting results
- getting the results used.

The Evaluation Design Process

Another part of front-end planning is to look ahead to the evaluation design process. The evaluation designs to be used will depend on the kinds of questions being asked, which will, in part, be based on the timing of the evaluation. Different designs will be appropriate in answering different kinds of questions. Figure 6.1 illustrates the evaluation design flow chart showing the process of preparing for evaluations and then conducting them.

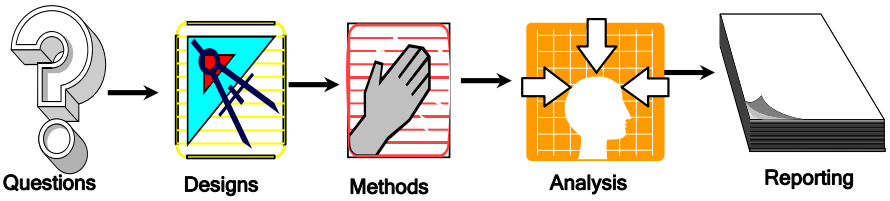


Fig. 6.1: Evaluation Process Flow Chart

The Evaluation Process

Ideally, the evaluation process begins with the initial program design. As the actual evaluation is developed, there are several distinct and important stages (see Figure 6.2).

Planning

- Get a thorough understanding of the program, project, or policy
  - Identify and meet with stakeholders
  - Explore context and gather background materials
  - Search for related, relevant evaluations
  - Meet with program staff
  - Review or develop theory of chain or logic model

Designing

- Determine the questions and issues
  - Meet with stakeholder
  - Determine general resources available
  - Assess stakeholders needs, including timing
- Prepare terms of reference and evaluation work plan
  - Identify type of evaluation
  - Identify specific evaluation questions
  - Select an appropriate design for each question
  - Select measures for each question
  - Develop data collection strategy, including instruments and sampling, if needed
  - Develop data analysis strategy
  - Review and test the methodology
    - Pre-test instruments
    - Train data collectors
  - Brief stakeholders on evaluation plan
  - Prepare work plan
  - Resource and time requirements **Doing**
- Gather the data
- Prepare data for analysis
  - Develop table shells
  - Clean data
- Analyze the data
- Formulate the findings

Reporting

- Identify major findings: what works, what does not
- Brief stakeholders on findings and statements of fact
- Develop recommendations: proposals for action
  - Clear and specific: who should do what and when?
  - Link evidence to recommendations

Feedback:

- Product: Written report, briefings
- Process: contributes to learning
- Involves stakeholders in developing a plan for implementing the recommendations (if any).

Fig. 6.2: The Evaluation Process

The **Planning** phase helps clarify the exact nature and scope of the evaluation. During this phase, many decisions are made and key issues identified: the purpose of the evaluation, the stakeholders, who will conduct the evaluation, how and when it will be conducted, the questions and issues, and how the results will be reported and disseminated. Typically, this is stated in the Terms of Reference (TOR). The needs of the sponsor of the evaluation, the key stakeholder will help shape the evaluation's nature and scope.

Terms of Reference (TOR) describe the overall evaluation and establish the initial agreements prior to the work plan. The process for developing the Terms of Reference can be very useful in ensuring that stakeholders are included in the discussion and decision-making about what evaluation issues will be addressed. It establishes the basic guidelines so everyone involved understands the expectations for the evaluation and the context in which the evaluation will both take place.

The heart of the evaluation planning is the **evaluation design**. It is the design that specifies the questions, the design for answering those questions, the necessary measures, data collection strategies including sampling, and data analysis techniques. If the design is flawed, it will limit the ability to make conclusions about the intervention. Generally, it is a good idea to present and discuss the methodology with the evaluation sponsor and other key stakeholders beforehand. An advisory group or peer reviewers may also serve as sounding boards.

The **doing phase** relates to the actual gathering and analysis of the data. Different kinds of data might be collected, so different analysis techniques are likely to be used. Based on the analysis, an initial set of findings is generated. These findings then need to be considered in terms of the context. The initial findings or statements of fact should be shared and discussed with the stakeholders.

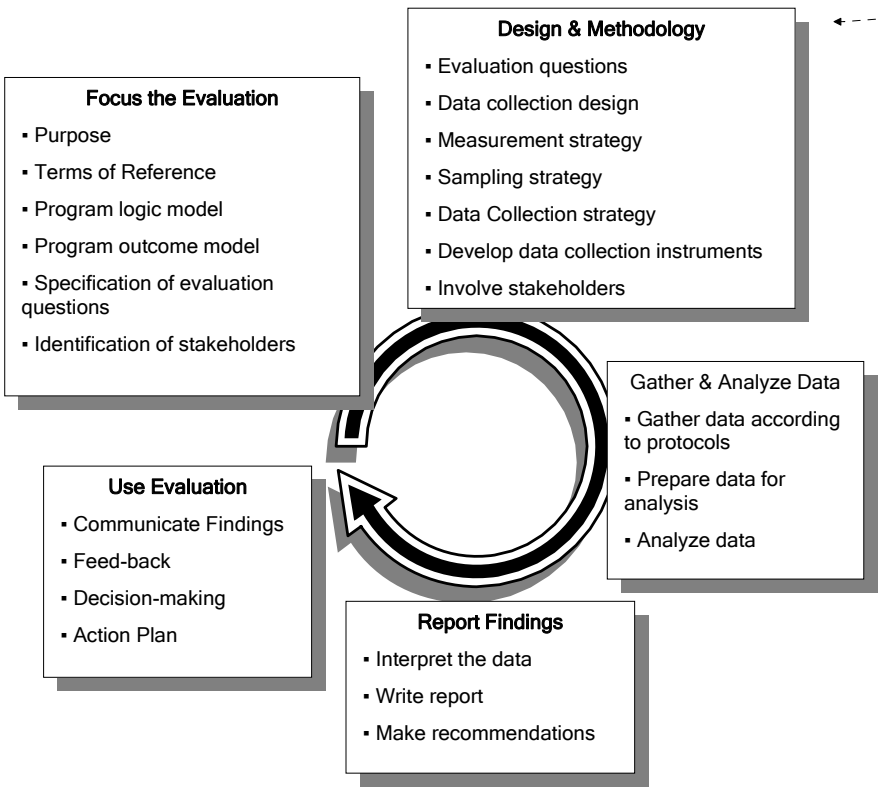
Once the analysis is completed, the results are written up, drafts are reviewed, and a final report is presented to the key stakeholders: the **reporting phase**. The report will typically provide background and context for the evaluation, the purpose of the evaluation, a description of the evaluation's scope and methodology, and findings (including both intended and unintended outcomes). The report may include information about lessons learned and/or recommendations. Understanding what works well and why it works well is as important as understanding what does not work well and why. The report should be written with the audience in mind; it should be clear and easy to read.

Major messages should be supported by evidence. Presenting data in charts or tables makes it easier for the reader to understand the major points. (Module 13 has more information about presenting results.) If recommendations are made, they should be clear, straightforward, and flow from the evidence presented. Ideally, the recommendations are specific in terms of who should do what. Not all reports come at the end of an evaluation or are in print. There are times when timely verbal presentations are needed to communicate unexpected or critical findings. Reporting plans should be flexible and designed with the client in mind.

Lastly, a good evaluation will include a **feedback process**. This means that while the report is in draft there is a review and dialogue with key stakeholders to discuss the findings and recommendations. A good evaluation is used; this may mean that a plan is developed to implement the recommendations. Many evaluations result in action to:

- modify intervention
- remove barriers
- inform future policy or interventions
- show others the way.

The relationship between the different components can be seen in Figure 6.3. The process typically is not linear; there is some back and forth between the elements.



Formatted: Bullets and Numbering

Fig.6.3: Approach to Development Evaluation

The Evaluation Design Matrix

An evaluation design matrix is another organizing tool to help you plan your evaluation. It is one we highly recommend. An evaluation design matrix organizes the evaluation questions and the plans for collecting information to answer the questions. The matrix, links descriptive, normative, and impact evaluation questions to design and methodologies. The evaluation design matrix is a very simple tool, but it can play a powerful role in planning and implementing an evaluation.

Which tool people use to help them think about a program, its context, measurable objectives and data collection and analysis strategies will vary. You may decide that you need to create your own. The point to all this is that you find a method that helps you to see all the pieces of the evaluation, to ensure they all connect, and that the connections are clear at every step.

The purpose of the matrix is to organize the evaluation purpose and questions and match what is to be evaluated with appropriate data collection techniques. Although there is no hard and fast rule, a design matrix usually includes the following elements:

- design evaluation questions and sub-questions
- indicators or measures
- data collection sources
- data collection methods/instruments
- sample.

Although each evaluation question is unique, data collection activities are designed to address more than one question and several activities may address a single question. The matrix incorporates known sources of information and planned sources. As you move from planning to implementation, sources can be expanded and clarified.

Beyond its immediate usefulness as a planning tool, the matrix can serve as a tool to promote collaboration between evaluators and program staff. During later stages of the evaluation, evaluators and program staff can review the matrix, update it and use it as a guide for implementing the evaluation.

Figure 6.4 is an example of an evaluation design matrix. Note that the matrix is worked horizontally. As you proceed through later modules, you will learn how to add information into this matrix.

Design Matrix Planning Instrument for:					
Questions:	Sub-Questions	Type of Question	Design	Measures or Indicators	Criteria for Normative Questions

Source: Revised from original presented by Linda Morra-Imas

Fig. 6.4: Design Matrix

Data Sources	Sample	Data Collection Instrument	Data Analysis	Comments

(Fig. 6.4: Design Matrix- – cont.

Connecting Questions to Design

You should now have your questions written. Your next challenge is to determine how you are going to find the answers. How do evaluation designs connect to the questions and why are they important? In and earlier module, we divided questions into three types: descriptive, normative and impact.

As you probably recall, descriptive questions describe the current situation (“what is”), normative questions compare "what is" to "what should be," and cause-effect questions determine "what difference the intervention makes."

To obtain good answers any question, you need to carefully plan and execute each step of the process. All evaluations pose their own special challenges in terms of gathering and analyzing data.

Answering Descriptive Questions

Descriptive questions, ask questions like “how many?” or “how much?” They may also ask for perceptions or opinions. To answer these questions, the common designs are:

- one shot
- cross-sectional
- before-and-after
- time series
- longitudinal
- case studies.

Descriptive questions generally use non-experimental designs.

These designs, when used to answer descriptive questions, do not involve a comparison group that did not receive the intervention. They focus only on those who have received the intervention.

One-Shot Designs

A **one-shot design** is a look at a group receiving an intervention at one point in time, following the treatment or intervention. We may use this design to answer questions such as:

- How many women were trained?
- How many participants received job counseling as well as vocational training?

We may use this design to ask participants how well they liked the training, for example, or to determine how they found out about the training: was it our outreach program or something else?

One-shot designs are represented as:

X O_1.

This means there is one group receiving the treatment "X" and we make one observation "O_1" sometime following the intervention. It is a one-time snapshot. There is no before-treatment or intervention measure.

Cross-sectional Designs

Cross-sectional designs also show a snapshot at one point in time. But in this case, we are interested in sub-group responses. This design is often used with a survey method. The subgroups may be based on age, gender, income education, ethnicity, or amount of intervention received.

Sometimes the evaluation question focuses on whether citizens or intervention beneficiaries are satisfied with services they received, or why they do not use services.

Sometimes the question may be to find out the current status of people who participated in an intervention a few years ago.

To answer these kinds of questions, a survey using a cross-sectional design might be used. A cross-sectional survey selects a sample of citizens, intervention beneficiaries, or former intervention participants at one in time. It would then gather data from them and report what they said.

A cross-sectional design might answer questions such as:

- Do participants with different levels of education have different views on the value of the training?
- Did women receive different training services than their male counterparts?

The cross-sectional design is represented as:

X O_1
O_2
O_3
"

That is, the observation is made after the intervention "X" and responses of subgroups ("O_1, O_2, O_3" and so on) receiving the intervention are examined.

For instance, in evaluating a program designed to economically empower women to launch their own small businesses, the evaluators may want to find out the views of women who have been through the program. Their views can shed light on whether what they learned in the economic empowerment program helped them launch a viable business, the kind of business they went into, and whether what they learned in program was useful for running the business. With limited resources, they may opt to conduct a short survey of recent program graduates (a one-shot design). Or they may survey both women who recent graduated from the program and women who had been through a traditional business skills program, or who had simply received seed money but no training or support aimed at economic empowerment (a static group comparison). These are non-experimental designs.

Before and After Designs

In this design, we ask about group characteristics before and after the intervention (also called pre- and post-intervention characteristics). There is no comparison group. Thus, we may ask whether our program participants increased their knowledge of parenting techniques, and test them at program entry and following program completion. We could look at the wages of our vocational training program participants before our training intervention and two years following the program.

Before-and-after designs are represented as:

O_1 X O_2.

That is, observation, intervention, observation.

Time Series Designs

Time series designs look for changes over time. The purpose, when used to answer descriptive questions, is to explore and describe changes over time either after, or before and after, the intervention. Thus, time series designs can be used to discern trends. They can be simple time series designs or cross-sectional designs. For example, child mortality rates might be examined before and after an intervention, providing maternal nutritional supplements, to identify the trends. Or changes in participant attitudes over time towards women entrepreneurs might be examined.

Simple time series is represented as:

$$O_1\ O_2\ O_3\ X\ O_4\ O_5\ O_6$$

where several (in this case, three) observations (often obtained from existing data) are made prior to the intervention and again three more after the intervention.

Longitudinal Study

A **longitudinal study** is another type of time series design. A longitudinal study is one in which repeated measures of the same variable are taken from the same people (or from sample groups in the same population). A panel design is a special type of longitudinal design in which a smaller group of people is tracked at multiple points in time, and their experiences are recorded in considerable detail.

Panel designs almost always use qualitative (open-ended survey questions, in-depth interviews, and observation) as well as quantitative data. Panel designs can give a more in-depth perspective on any changes people may be experiencing as a result of an intervention. For example, a study looking at Poland's family allowance used panel data gathered from the same people between 1993-1996 to find out how the allocation of social benefits affected families' transition into and out of poverty.

Case Study Designs

The descriptive case study is also a design. In-depth information is collected over time to better understand the particular case or cases. This is an often-used design for development interventions. The case study design is particularly useful for describing what implementation of the intervention looked like on the ground and why things happened the way they did. The descriptive case study may be used to examine program extremes, or a typical intervention. (More information on case studies is at a later section in this module.)

Case study design is represented as:

O_1
O_2
O_3

Answering Normative Questions

What designs work best for normative questions? The logic for normative questions is similar to descriptive questions, except that normative questions are always assessed against a criterion. The difference between answering normative questions and descriptive questions is only that there is a specified desired or mandatory goal, target, or standard to be reached and the actual findings are compared to that standard. Generally, the same designs work for normative questions as descriptive questions.

Answering Cause-Effect Questions

Cause-effect questions pose the greatest challenge; this is where we need a well-thought out design (as opposed to a particular approach to answering a relatively straightforward descriptive question). In any evaluation with questions involving impact, the evaluation design attempts to rule out feasible explanations for the observed results – other than the intervention – in order to conclude that the invention had an impact. In other words, we try to be sure that any observed changes can be attributed to the intervention, rather than something else.

When we talk about eliminating other possible explanations, we are talking about **internal validity**. Internal validity refers to the design's ability to rule out other explanations for the observed results. An evaluation design with strong internal validity enables you to be much more confident that the intervention caused the observed results. A design with weak internal validity makes it harder to convince others that the intervention caused the observed results. Yet, it is important to keep in mind that these threats are just possible rival explanations; they might not actually exist. Thus, internal validity is very context-related.

Common threats to internal validity are:

- history (events occurring at the same time)
- maturation of subjects (just getting older changes the results)
- testing (learning how to take the test)
- instrumentation (changes in data collection instruments or procedures)
- regression to the mean (averaging out of scores over time)
- selection bias (participants may be different to begin with)
- attrition (a specific group of people may drop out).

Experimental and quasi-experimental designs have been used to control for these threats to internal validity. Experimental designs are sometimes called the "medical model." They involve randomly assigning program participants to a group to maximize the probability that the groups are equal on factors that could influence the program or intervention results – like age, gender, education, attitudes, past history, and so forth. Think of a new drug treatment where patients are randomly assigned to the new drug, current alternate drug therapies, or a placebo. Patients, and the health-care providers, do not know who is getting which drug in these **single- or double-blind studies.**

In the development evaluation field, many believe that they cannot use these experimental evaluation designs because it is unethical to withhold the "best" alternative from a group of participants, or because political decisions about who gets the intervention prevent randomization. The Word Bank has shown, however that randomization is more frequently possible then perhaps thought, since there are often not enough resources for an intervention to be delivered to all eligible participants initially.[1] A new textbook and curriculum may be introduced only in some districts initially. But the key is randomly assigning districts to intervention and non-intervention groups so as to systematically test for eventual differences in academic performance.

1 Robert Boruch. Ethics and Randomized Trials. Presentation made at the International Program for Development Evaluation Training (IPDET) – 2004.

To reduce the possibility of being fooled into thinking we know something as true, which really is not, we can borrow from social science methods. Using experimentation, it is possible to control the implementation of a program, policy, or project, who receives it, and the environment in which it is delivered. When you can control everything but the intervention, you can then be fairly certain that any differences you observe are the result of that intervention.

Suppose you have a fertilizer that is intended to increase the crop yield of corn. Let us say you have a greenhouse where you can control the temperature, water, and soil conditions. You create two separate growing areas within the greenhouse and randomly plant the seeds in the testing area and the control area. Both areas receive the same temperature, sunlight, and water and are planted in the exact same soil mixture. The test area receives the fertilizer and the control area does not. At harvest, the yields are measured. If the test area has a higher yield than the control area, you will conclude the fertilizer made a difference.

Now let us think about what happens when you are working in the field instead of the controllable environment of the greenhouse. It is possible that your two areas are close together and that fertilizer can run off into the non-test area, thus giving you an imperfect measure of the impact of the fertilizer. You might move the non-test area to a different part of your field but the soil, light, temperature, or rain may be slightly different. The two fields may also receive different amounts of attention. While you can still measure impact, you may be less certain about concluding that the fertilizer alone caused those results.

In the complex world in which development interventions take place, it becomes difficult to identify impact in the midst of other factors. In the agricultural case, let us suppose an irrigation intervention was implemented during a time of ideal weather and strong demand for the crops. The income in the area where the irrigation intervention was implemented increased over prior years. But is the higher income *a result* of the intervention? Or is it caused by other factors, such as increased rainfall, general economic good times, or an unusual period of political stability.

Multi-site and cluster evaluations present special challenges as well as unique opportunities for looking at cause-effect questions. The context at each site and the nature of the intervention often vary considerably across different locations. This can help identify some of the situational and implementation factors that enhance or hinder impact – but the analysis can be considerably more complex. This complexity may limit the extent to which the data collection and analysis phases can be conducted in a highly participatory mode, and it may also limit the viability of rapid assessments as an option.

Evaluation designs can help us determine the impact of a program to the extent that they give us control over the implementation and measurement of the program. The intent is to eliminate other possible explanations for what we observe.

Design Elements

In this module, we will focus first on evaluation designs that address cause-effect questions and then discuss designs for addressing descriptive and normative questions. We will postpone discussion of data collection and analysis methods, reporting, and utilization of results – all elements of a complete evaluation design – since these topics will each be individually discussed in later modules.

Design Elements for Cause-Effect Evaluations

To answer whether an intervention made a difference, the evaluation has to show that the key measures changed as a result of the intervention. The following are some of the elements to consider in cause-effect evaluation design:

- before-and-after measures
- comparison groups
- random assignment to the comparison groups
- use of control variables
- use of natural variation
- causal tracing strategies.

Evaluators doing *traditional* experimental evaluation will focus on the first three of these elements of design:

- before-and-after measures
- comparison groups
- random assignment to the comparison groups.

For some of the *newer approaches* (such as cluster and multi-site evaluation and rapid assessment), evaluators may draw on other tools, including:

- use of control variables
- use of natural variations
- causal tracing strategies.

Before-and-after Measures

Before-and-after measurement is one way to measure change. Change is measured by comparing key measures after the intervention began against the measures taken before the intervention began. Before and after measures are also called pre- and post-tests.

The "before" measure might be called a **baseline**. Many evaluations collect before data, called baseline data. Collecting baseline data is sometimes called a baseline study.

However, a design with only before-and-after measures is insufficient to demonstrate that the intervention alone *caused* the change. Maybe the people changed their behavior because they were being observed or maybe something else occurred at the same time as the intervention and that something was the real cause of the changes we observed.

It is also important to remember that in a situation where there is little change in the measures, you should be hesitant to conclude that the intervention did not work.

For example, an intervention (the treatment) to reduce poverty was implemented in a country. Everyone was eligible to receive the intervention so there was no comparison group. At the end of ten years, the proportion of people in poverty did not change. Can you conclude that the poverty reduction intervention did not work? It could be that without the intervention, a greater proportion of people would have been in poverty. A common way to determine if an intervention had the desired effect is to use comparison groups. You compare results from the group that did receive the treatment to a group that did not receive the treatment.

Comparison Groups

Without a **comparison group**, you cannot tell if poverty would have been worse if the poverty reduction intervention was not in place. The intervention may have been effective in holding the line on poverty in spite of declining economic conditions.

To strengthen this design, you would want to compare your intervention group to a group who did not receive the intervention.

The group that receives the treatment may be called the **treatment group**.

The group that does NOT receive the treatment may be called the **control group**.

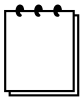
If an intervention causes change, then those who participate in the intervention will show more change than those who did not.

Returning to the example of the irrigation intervention, we would need to compare a similar agricultural area that did not receive the irrigation program with the one that did. If the income in the area that received the irrigation program increased while the income of the comparison group did not, then we would have stronger evidence that the irrigation program caused a positive outcome.

However, not all the other factors have been ruled out. What if the community that receives the irrigation intervention is different in some important ways from the comparison community that did not? Maybe other industries moved into the intervention area, thus increasing income. Or maybe the comparison community grew different crops and their crops did not command as high a price as the crops grown in the irrigation community.

In some interventions, there will be those who receive it and those who do not. This design can be used in evaluation, as might be done in the evaluation of a job training program for laid-off workers (see Case 6-1).

Case 6-1: Impact of Job Training Programs for Laid-Off Workers



Many developing countries face the problem of retraining workers when state-owned enterprises are downsized. A training program was put into place and subsequently evaluated. Evaluating training programs are challenging because they frequently have several different components to serve different constituencies, and there are variety of ways to measure outcomes: employment, self-employment, monthly earnings, and hourly earnings.

Questions: Are participants more successful in re-entering the labor market than are non-participants? What is the cost effectiveness of the different training programs?

Design: Participants receiving training were matched with a similar group of non-participants. Administrative data and survey data were used. A variety of statistical techniques were used to measure the impact of the training program.

However, it is possible that the people who participated were different in some way that made a difference in terms of their outcomes. Maybe they were more motivated or had more job experience, so it was easier for them to find new employment. To strengthen this design, it is necessary to randomly assign people to participate or not to participate in the program.

Random Assignment

In the ideal world of science, we would be able to randomly decide who or what will receive (or not receive) the intervention. Random assignment means that people, things, farms, roads, cases, or whatever are placed in their groups by chance. **Random assignment** makes the groups comparable. For example, if we randomly assign communities to receive the irrigation intervention or not, we make the groups relatively equal in terms of a variety of factors we cannot control. Given the equality of the groups in terms of a host of factors, we would feel more confident in concluding the intervention had an impact if the intervention group had higher income than the control groups.

However, random assignment is not always an option. Sometimes all eligible people are supposed to receive the intervention, and/or it would be unethical to withhold it from some of them. Sometimes you cannot make people participate in an intervention. In this instance, it may be more practical to provide an intervention in a random selection of sites rather than provide interventions in multiple sites and then randomly assign participants.

It is possible, however, to be in a situation where the intervention is not large enough to accommodate all those who apply. A project manager might want to assign those with the best chance of benefiting from the intervention to participate. This may be a way to get the most benefit from limited program dollars. For example, the training program is likely to show the best results if it chooses highly motivated people to participate. However, from an evaluation perspective, if the best people are assigned to the program, there is likely to be a bias in the results. Actually, in this type of situation, random assignment may not only enable the use of a strong design to measure impact, it may also be more equitable; no bias and favoritism is in play when assignment is based on chance.

Use of Control Variables

As noted above, sometimes it is not possible to randomly assign people to groups for comparison purposes. However, it might be possible to find a group of people (or a village, etc.) who will not be receiving the intervention yet, and who could be used to compare results with. But without random assignment, there is still a good chance that the comparison group differs in important ways from the group receiving the intervention. If that is true, the intervention might look effective (or ineffective) simply because of pre-existing differences between recipients and non-recipients.

In such cases, one option is to collect data about factors that might be different between the two groups, and that seem likely to impact outcomes. These variables are then built into the data analysis as **control variables**. Using control variables allows us to rule out some alternative explanations even when random assignment is not possible.

Use of Natural Variation

Sometimes it is not possible to find a meaningful comparison group of any description, especially if the intervention has been implemented across a wide area, as is the case with multi-site and cluster evaluations. Does this mean that there is no hope of figuring out whether any changes observed are due to the intervention or not? Not at all; and the answer lies in **natural variations** in the intervention.

Suppose a prenatal care program has been implemented in all the villages in a particular geographical area, and you have been asked to conduct a multi-site evaluation. As with any large intervention, more often than not there is some variation in the extent to which the program actually reaches its target population, and also in the quality of the services delivered. Now, suppose you find better health in newborns and lower infant and mother mortality in those villages where more mothers received more frequent and better quality service. Such evidence adds to the confidence with which you can attribute health outcomes to the intervention, especially compared to looking only at overall outcomes across all participants.

Causal Tracing Strategies

Many of the strategies described above for determining whether observed changes are due to the intervention (as opposed to some other cause) require a fairly structured and heavily quantitative data collection strategy. They will also require a relatively complex statistical analysis. For the evaluator who is conducting a rapid assessment, or one who is evaluating a very small, or new and untried intervention, such strategies may be neither practical nor advisable.

What options are available when the sample size is small, the data collection strategies are largely open-ended, and/or when sophisticated statistical analysis is not feasible? There are some resources available on how to use **causal tracing strategies**, which are particularly suitable for qualitative and mixed method (qualitative and quantitative) studies, and especially when the target population is small.

Broadly speaking, the principle is the same – systematically rule out alternative explanations until you are as sure as you need to be that the changes you observed are indeed caused (primarily or at least substantially) by the intervention – or are not.

The following is a list of nine possible sources of evidence you might gather when using causal tracing (adapted from Davidson, 2001):

1. **Causal list inference**:
   - we know that this particular outcome is almost always caused by one of the following: A, B, C, or D, and on this occasion neither B, C, or D occurred, so we can be almost sure the cause was A.
2. **Modus operandi (MO) inference** – use this if more than one possible cause occurred:
   - we know that this outcome is almost always caused by one of the following: A, B, C, or D, and on this occasion, neither C nor D occurred, which narrows the cause down to A or B. In addition, only the characteristic causal chain/MO/telltale pattern of events for A was present
   - this inference is strengthened if the MO for A is highly distinctive/very different from that for B.
3. **Temporal precedence** – the observed effect only happened after the intervention had begun, not before.
4. **Constant conjunction** – the effect was observed everywhere the intervention was implemented.
5. **Contiguity of influence** – a plausible mechanism links the intervention with the observed effect
   - this inference strengthened if we can find evidence of the mechanism in action.
6. **Strength of association** – the observed change was much stronger where the program was implemented than it was where other possible causes were present.
7. **Biological gradient** – the more treatment someone/a village received, the larger the observed change.
8. **Coherence** – the relationship we see between the intervention and the observed change fits logically with other things we know about the intervention and this particular outcome.
9. **Analogy** – the pattern we see between the intervention and the observed changes resemble the well-established pattern we know about between a related intervention and its effects.

When designing your data collection strategy, you should consider which of the above pieces of evidence it is feasible and necessary to gather, and plan ahead how you will obtain them. You certainly do not need all of these to be able to make causal attributions; just gather the pieces that make the most sense, and that together will give you sufficient certainty about your findings, given the decisions that will be based on your evaluation.

Types of Designs for Cause-Effect Evaluations

When addressing cause-effect questions, evaluation design needs to address the question "What would the situation have been if the intervention had not taken place?" It is not possible to actually observe this but it is possible to estimate what might have happened. You can try to construct an approximate group of non-participants in a program.

In order to evaluate impact, you might consider one or more types of evaluation designs. They are:

- experimental design
- quasi-experimental design
- correlational design
- case study design
- non-experimental design.

Experimental Design

The classic **experimental design**, sometimes called the true experiment, is considered the strongest design for cause-effect questions. Experimental design rules out most other possible explanations. The following are characteristics of experimental design.

- Random assignment is the essential component of experimental design. Random assignment assures that the comparison groups are comparable.
- One group is randomly assigned to receive the intervention and another randomly assigned group serves as the control group and does NOT receive the intervention
- Experimental designs usually contain before-and-after measures for the comparison groups. However, they may also use "after" measures only. In these cases, there is no "before" measure as the two groups are assumed to be equivalent before treatment because of the random assignment.

The main benefit of experimental design is that it is easier to draw conclusions from the results because it better controls threats to internal validity. With random assignment, the probability is increased that people in the treatment and control groups are the balanced or equivalent in characteristics such as interest in the program, past history, current events, attitudes, and the like.

If the group that received the intervention shows a change and the control group does not, you can draw conclusions on the intervention.

While experimental designs are considered the optimum approach for evaluating an intervention, they can be difficult to implement for several reasons:

- Randomization may be unethical owing to the denial of benefits or services to otherwise eligible members of the population for the purposes of the study.
- It can be politically difficult to provide an intervention to one group and not another.
- The scope of the intervention may rule out the possibility of selecting a control group such as with a nationwide program or policy change.
- Individuals in treatment or control groups may change certain identifying characteristics during the experiment that could invalidate or contaminate the results. For example, people who were denied a program benefit may seek it through alternative sources, or those being offered a program may not accept the intervention.
- It may be difficult to ensure that assignment is truly random. An example of this might be administrators who exclude high-risk applicants to achieve better results.
- Experimental designs can be expensive and time consuming in certain situations, particularly in the collection of new data[2].

One way to address some of these issues with experimental design is to bring the control group into the experimental group at a later stage, when beneficial results have been demonstrated. In this way, the random selection determines *when* the person receives the intervention, not *if* they receive the intervention.

---

[2] The World Bank, Poverty Net, Evaluation Designs, online site: http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:20188242~menuPK:412148~pagePK:148956~piPK:216618~theSitePK:384329,00.html

Patton[3] identifies at least six conditions for which randomized control trials (RCTs) are especially appropriate. Consider RCTs when:

- you have a discrete, concrete intervention (treatment): that is, singular and well-specified
- implementation can be standardized
- implementation can be controlled
- valid and reliable measures exist for the outcome to be tested
- random assignment is possible
- random assignment is ethical.

The evaluation of the impact of textbooks in a developing country is an example of an experimental design, with random assignment. (See Case 6-2.)

Case 6-2: Textbooks and Test Scores

Prior studies have suggested that the provision of textbooks is a cost-effective way of increasing test scores. The main question is: What is the effect of textbooks on learning outcomes? Learning outcomes are defined as test scores.

Design: All grades in a randomly selected subset of 25 out of 100 rural primary schools were provided with textbooks. Achievement tests were given to the students before the textbooks were distributed, and then again 10 months later. The same tests were also given to the students in control schools. The analysis compares differences in the pre- and post-test scores between the control and treatment schools.

Comments: Clearly, there are ethical issues here with regard to withholding textbooks from the 75 schools that didn't receive them. However, as mentioned before, quite often funding is not initially available to supply a product, service, or intervention to all potential recipients. Here, if the textbooks are found to have an effect on learning outcomes, this helps justify the expansion of the program (and its funding) to later provide textbooks to the other schools as well.

[3] Michael Q. Patton. The debate about randomized controls in evaluation: The gold standard question. PowerPoint presentation to IPEDET, June 2005, slide 18.

Patton[4] identifies examples where he feels RCTs are NOT appropriate. They are for:

- complex, multi-dimensional and highly context-specific community interventions
- ethical constraints.

He offers one statement that shows this clearly:

"No randomized control trials of parachute use have been undertaken."

Patton also identifies the following where he feels the use of RCTs is not needed:

- face validity is high
- observed changes are dramatic
- link between treatment and outcome is direct.

The Campbell Collaboration (C2) is an international organization that is fostering and showcasing the use of experimental designs in evaluation.

The Campbell Collaboration is a non-profit organization. Its aim is to help people make well-informed decisions about the effects of interventions in the social, behavioural, and educational arenas. The objective of the C2 is to prepare, maintain, and disseminate systematic reviews of research on interventions. They acquire and promote access to information about trials and studies of interventions. C2 also builds summaries and publishes electronic brochures of reviews and reports of trials for use by policy makers, practitioners, researchers, and the public.

You might find it helpful to search their library at their website to read about evaluations and the designs they used. Their website is at: http://www.campbellcollaboration.org/

[4]Michael Q. Patton, The debate about randomized controls in evaluation: The gold standard question. PowerPoint presentation to IPEDET, June 2005, slides 20 – 30.

Quasi-Experimental Designs

Quasi-experimental design means that the design is similar to true experiments except that the comparison groups have NOT been formed by random assignment. Quasi-experimental design methods can be used to carry out evaluation when it is not possible to construct treatment and control groups using random assignment.

In quasi-experimental design, you construct groups that are as equivalent as possible. You want to compare groups that resemble the treatment group. Sometimes you can create a comparison group by matching key characteristics. Other times you will find a comparison that is not exactly the same as the group that received the intervention, but is similar enough to provide some comparison.

For example, you might be able to compare a village that received an economic development intervention with one in the same geographic area that did not. Often, you have the specific group that received the intervention and are looking for an appropriate comparison group. In these designs, you may have to collect more information to make a case that the intervention outcomes cannot be explained by factors other than the intervention.

Common examples of quasi-experimental designs are:

- before-and-after – good for descriptive questions
- matched and non-equivalent comparison design
- time series and interrupted time series design
- correlational design using statistical controls
- longitudinal design
- panel design.

An example of a quasi-experimental design is the evaluation of El Salvador's EDUCO Program.[5] (See Case 6-3.)

Some of the same basic designs used to answer descriptive questions can be used to answer cause-and-effect questions. The key difference is that for cause and effect questions, comparison groups are needed, either using random assignment (experimental design) or constructed comparisons (quasi-experimental designs).

---

[5] The World Bank, Development Research Group. "Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program." Impact Evaluation of Education Reforms, Paper No. 8. February 1998.

Case 6-3: Do Community-Managed Schools Work?

An Evaluation of El Salvador's EDUCO Program

This evaluation was intended to measure the effects of decentralizing educational responsibility to communities and schools on student outcomes. El Salvador's Community-Managed Schools Program (EDUCO) was designed to expand rural education rapidly following a civil war. It compares student achievement on standardized tests and school attendance of rural students in EDUCO schools to those in traditional schools. It controls for student characteristics and selection bias using statistical controls.

In 1991, the Minister of Education expanded education in rural areas through the EDUCO program. This is an innovative program for both pre-primary and primary education to decentralize education by strengthening direct involvement and participation of parents and community groups. An elected Education Association drawn from the parents of the students manages the EDUCAO schools.

The question is whether quick expansion to rural area has come at the expense of learning. This study compares outcomes measures of 3rd graders in EDUCO and traditional schools. Outcome measures are based on standardized tests in mathematics and language. However, because tests scores may be unresponsive in the short term, the evaluators also looked at the school days missed by a student due to teacher absence.

Differences in educational outcomes, however, can be affected by factors other than school. These include differences in household background, the school's inputs, and organizational factors.

The evaluators needed to determine whether differences in test scores (as a measure of student achievement) were due to differences in type of school, or to other factors. Factors apart from type of school (EDUCO or traditional) that might explain student achievement are:

- Household characteristics (education, family size, income)
- Student characteristics (gender, age, number of siblings)
- School data (enrollment, teacher quality, school facilities and finances)
- Teacher characteristics (educational background, years of experience).

The evaluators used data collected by surveys administered by the Ministry of Education to construct a model that would measure the independent impact of the type of school while controlling for those other factors. Using complex statistical modeling that controlled for all of the above factors (see earlier discussion of control variables), the evaluators concluded that the achievement scores of children in EDUCO and traditional schools are about the same. The rapid expansion did not have an adverse impact on learning, even controlling for a range of other variables. In other words, the community-managed schools were as effective as regular schools.

Correlational Design

Correlational designs are often used when we are seeking to answer questions about relationships. Correlational designs can be used with data already available, or with new data. For example, if you wanted to find out if having women in political offices is related to a more honest government, a correlational design could be used. Data about the proportion of women in political office in the different areas within a country and the amount of reported corruption could be correlated with each other.

Correlational evidence alone cannot establish causality – even if governments with more women in office are correlated with less corruption, it would still be necessary to rule out any plausible alternative explanations for the relationship.

Case Study Design

A case study design is frequently used when the researcher wants to gain an in-depth understanding of a process, event or situation. It is useful when the question deals with how something works or why something happens, and is especially useful when the intervention is relatively innovative or experimental or not well understood. Case studies are frequently used in evaluating development interventions.

Case studies can use qualitative and/or quantitative methods to collect data. They can consist of a single case (a one-shot, before and after or single time-series design) or they can consist of multiple cases (comparative designs). They can focus on in-depth understandings of individuals, an organizations, communities, programs, cities, and/or nations.

If we were interested in monitoring public transportation in a country, we might simply track key indicators against the baseline and targets. We might do a national study if the indicators are the number of miles covered by public transportation, capacity, the number of people who use the system, and revenues received. However, if we wanted to answer other kinds of questions that require more in-depth data collection, we would opt for case study. We might choose a single case study or we might choose a few different locations to gain a wide range of experiences.

For instance, if we wanted to learn more about people's choices about using public transportation, we will have to gather the data from them. This will require more resources to collect this data if done on a national scale. It is more manageable to gather that data in a more narrowly defined geographic area – a single case. Alternatively, researchers might opt for a multiple case study, where several cities might be selected.

The cases might be selected in several ways:

- randomly
- judgmentally or purposively selected based on some specific criteria
  - best case, typical case, worst case, or a mix
  - only large cities or cities of varying sizes.

The same data collection strategies used in the single case study can be used in multiple case studies.

Case studies make sense in development where the intention is about understanding a specific situation in order to make or adjust policy or practice. But not only are case studies more practical than trying to do large national studies, they also provide in-depth information that is often appropriate to the decision-maker's focus. A comparative case study of the use of free immunization clinics would provide greater understanding about why one is more successful.

Key Points about Design

Each evaluation will be different. Each will have different evaluation questions, data available, time constraints, and limitations of resources in the country. As an evaluator, you must explore the options for your design in an attempt to give you the most robust results.

The following are key points to keep in mind as you design your evaluations:

- There is no perfect design.
- Each design has strengths and weaknesses.
- There are always trade-offs in terms of time, cost, and practicality.
- Acknowledge trade-offs and potential weaknesses.
- Provide some assessment of their likely impact on your results and conclusions.

Michael Scriven's[6] *Key Evaluation Checklist* offers the following examples of questions that have to be answered in the design phase:

- Do you have adequate domain expertise? If not, how will you add it to the evaluation team (via consultant(s), advisory panel, full-team membership)?
- Can you use control or comparison groups to determine causation of supposed effects?
- If there is to be a control group, can you randomly allocate subjects to it?
- Can you double- or single-blind the study?
- If a sample is to be used, how will it be selected, and if stratified, how will it be stratified?
- If none of these, how will you determine causation (of effects by the evaluand; depending on the task, you may also need to determine the contribution to the effects of various components of the evaluand)?
- Will/should the evaluation be goal-based or goal-free?
- If judges are to be involved, what reliability and bias controls will you need (for credibility as well as validity)?
- How will you search for [anticipated and unanticipated positive and negative) side-effects?

On the following pages you will find several tables that summarize information in this module. You might find them helpful to print out and use for reference.

[6] Michael Scriven,. *Key evaluation checklist*, October 12, 2005. http://www.wmich.edu/evalctr/checklists/kec_october05.pdf

Linkages between Your Question and the Design

Descriptive Questions:	Non-experimental, quasi-experimental, or qualitative approaches
Normative Questions:	Non-experimental, quasi-experimental, or qualitative approaches Plus goals/standards/needs assessment
Cause-Effect Questions:	Experimental, quasi-experimental, or non-experimental with in-depth causal tracing

Summary of Common Designs for Data Collection

Experimental designs:	Always use random assignment to treatment and control groups. True experiment collects data before and after treatment. Variations: sometimes only collect data after treatment.
Quasi-experimental designs:	Compares intervention and non-intervention groups: no random assignment.
Matched:	Where the groups are matched on key characteristics.
Non-equivalent groups:	Comparison of group with intervention to group without the intervention.
Correlational design:	Collects data from all or a sample of people, cases, units (etc.), and uses statistical techniques to determine whether there are relationships.
Cross-sectional design:	Collects variables from a sample of cases or people at one point in time. Uses statistical controls to separate cases into those who received the intervention and those who did not.
Interrupted time series:	Collects the same data at many points in time before and after the intervention from different people or the same people.
Longitudinal design:	Collects the same data at a few points in time from the same people or from different samples of people from the same population.
Panel design:	Collects in-depth qualitative and quantitative data from the same people at various points in time.
Non-Experimental designs:	Designs for descriptive questions
Cross-sectional design:	Collects variables from a sample of cases or people at one point in time
Time-series design:	Collects the same data over time, before and after an intervention to observe trends
Descriptive case studies:	In-depth information across few sites
Before-and-after design:	Collects data on key measures before and after intervention
One Shot:	A snapshot – no before measures and no comparison

IPDET Terms	Design Types	Visual Representation	Key Advantages	Key Disadvantages
Experimental	**Experimental designs are characterized by random assignment to control and intervention groups.**			
	Randomized Comparison group	0_1 X 0_3 0_2 0_4	strong internal validity, identifies change over time both with and without intervention	costly, ethical considerations, difficult to generalize
	After only *Randomized* Comparison group No before test	X 0_3 0_4	good internal validity, slightly more practical, useful for comparing outcome	does not identify change over time
Quasi-experimental	**All quasi- pre-, and non-experimental designs are slightly weaker than experimental designs with respect to validity or have a low internal validity. Quasi-experimental designs involve comparisons but without random assignments**			
Matched and non-equivalent	Before & after Between Groups (non-equivalent) comparison	0_1 X 0_3 0_2 0_4	context must be considered, greater confidence than within group comparison	rules out the effect of history, difficult to control for all variables which make groups non-equivalent
Time Series Interrupted Time Series (good for descriptive questions)	Time Series (within group)	$0_1\ 0_2\ 0_3$ X $0_4 0_5\ 0_6$	threat of history partially controlled, maturation controlled	threat of testing bias
	Time Series Between Groups (non-equivalent) Comparison	$0_1 0_3\ 0_5$ $0_7 0_9 0_{11}$ X $0_2\ 0_4\ 0_6$ $0_8 0_{10}\ 0_{12}$	rules out threats of history, regression toward mean reduced	costly, time consuming, difficult to keep track of people over time
Correlational		0_1 0_2 0_3	uses statistics to determine correlations between cases to isolate potential threats determines important relationships and potentially confounding variables	requires large sample sizes, no statement about cause can be made, speculative
	After only with Non-equivalent Comparison Group	X 0_1 0_2	practical, context must be considered, control of effects of testing, instrumentation, regression, history	ethical considerations, selection threatens validity
	After only with Different Treatments Design	X 0_1 Y 0_2 Z 0_3	can compare interventions, must take context into consideration	many threats remain
Longitudinal	No Baseline	X 0_1	follows individuals over time	costly, difficult to keep track of individuals over time
Panel	Measure same over time	$0_1\ 0_2\ 0_3\ 0_4\ 0_5\ 0_6$	in depth information	can be costly
Non-experimental	**Ideal for Description**			
One shot		X 0_1	ease, practicality	many threats to validity, weak design
Before and After (good for descriptive questions)	Within Group Before & after Design	0_1 X 0_3	practical, context must be considered	testing, instrumentation and regression threats
Cross Sectional	Within and Between	X 0_1 0_2 0_3	clear picture of a point in time	no clear indication of what is happening over time
Case Study		0_1 0_2 0_3	in depth contextual information	time consuming, little internal validity

Adapted from Grembowski, D. (2001). *The Practice of Health Program Evaluation.* London: Sage Publications. Adapted for IPDET June 21, 2004.

No Perfect Design

Experimental Design:

- Controls for internal threats to validity
- Hard to do in the public sector.

Before and After Design:

- Useful in giving context for measuring change.
- Depending on the situation, it may have some weaknesses: testing, instrumentation, regression to the mean, attrition, history, and maturation may be threats.

Comparison Design:

- Useful in looking at differences between groups
- Controls for history and maturation if comparison group is a close match.
- Selection and attrition are threats.

One Shot Design:

- Useful for descriptive and normative questions
- Very weak for cause/effect questions: many threats
- Multiple one-shot designs begin to build a case.

Summary

In this module you learned about evaluation design. Check each of the following that you are able to complete. Review any topics that you cannot.

- ❑ define evaluation design
- ❑ describe how evaluation designs connect to evaluation questions
- ❑ describe elements of cause-effect evaluation design, including: before-and-after measures, comparison groups, random assignment, use of control variables, use of natural variation, and causal tracing strategies.
- ❑ list and describe the causal tracing strategies
- ❑ describe the types of cause-effect evaluation design, including: experimental design, quasi-experimental designs, and non-experimental design
- ❑ list and describe the kinds of quasi-experimental design, including: matched, non-equivalent groups, correlational, cross-sectional, interrupted time series, longitudinal, and panel design
- ❑ list and describe the types of non-experimental design, including: cross-sectional, time series, descriptive case studies, before and after, and one-shot.

Quiz Yourself

Answer the following multiple-choice questions to help test your knowledge of evaluation design.

You will find the answers to the questions on the last page of this module.

1. Which of the following shows the correct **sequence** for the **process of evaluation design**?
   a designs, questions, methods, analysis, reporting
   b. analysis, designs, questions, methods, reporting
   c. methods, questions, designs, reporting, analysis
   d. questions, designs, methods, analysis, reporting
2. Which of the following lists the **phases of the evaluation process in the correct order**?
   a. doing, planning, reporting, feedback
   b. planning, doing, feedback, reporting
   c. doing, reporting, planning feedback,
   d. planning, doing, reporting, feedback
3. If you are doing a **traditional experimental evaluation**, you will likely focus on which of the following?
   a. causal tracing strategies, random assignment
   b. use of natural variable, use of control variables
   c. use of control variables, causal tracing strategies
   d. before and after measures, comparison groups
4. What is a **baseline**?
   a. the "before" measure
   b. the "after" measure
   c. the "research" measure
   d. the "during" measure
5. In **before-and-after measures**, how is measurement done?
   a. measure the difference between key measures of two comparison groups
   b. measure the difference between key measures of one group before the intervention and after the intervention
   c. measure the differences among random groups on key measures

6. Which of the following is a description of **causal list inference**?
   a. a particular outcome is almost always caused by one of A, B, C, or D, and on this occasion neither B, C, or D occurred, so we can be fairly sure that the cause was A
   b. more than one possible cause occurred; this outcome is almost always caused by A, B, C, or D, and on this occasion, neither C nor D occurred, which narrows down the cause to A or B. In addition, only the characteristic causal chain/MO/telltale pattern of events of A was present – this inference is strengthened if the MO for A is highly distinctive/very different from that for B
   c. the relationship we see between the intervention and the observed change fits logically with other things we know about the intervention and this particular outcome
   d. a plausible mechanism links the intervention with the observed effect – this inference is strengthened if we can find evidence of the mechanism in action
7. Which of the following is a description of **coherence**?
   a. the observed effect only happened after the intervention had begun, not before
   b. the effect was observed everywhere the intervention was implemented
   c. the observed change was much stronger where the program was implemented than it was where other possible causes were present
   d. the relationship between the intervention and the observed change fits logically with other things we know about the intervention and this particular outcome
8. Which of the following gives a list of examples of **quasi-experimental design**?
   a. panel design, longitudinal design, one shot design
   b. time series design, before and after design, descriptive case design
   c. matched, non-equivalent groups, interrupted time series
   d. cross-sectional design, correlational design, descriptive case study

9. A **descriptive case study** is often used for which of the following reasons?
   a. repeated measures of the same variable are taken from the same people
   b. tracking key performance indicators over time before and after an intervention
   c. seeking an answer to questions about relationships between or among data
   d. researcher wants to gain an in-depth understanding of a process, event, or situation
10. List five key points about design.

Reflection

Think about the kinds of evaluation that you are asked or may soon be asked to complete.

- What kinds of evaluation designs have you used in the past?
- What did you like about evaluation designs that you have chosen in the past? What did you dislike?
- Knowing what you have learned now, what would you change about past evaluation designs.
- What evaluation designs are you considering for future evaluations? Why? What are the key considerations for your choice of this (these) designs?

Ev

Hints for Development Evaluators

- Choose your evaluation design based upon the evaluation questions you wrote earlier.
- There is NO perfect design, each has strengths and weaknesses.
- Choose your design based on trade-offs of time, cost, and practicality.

Mgr

Hints for Development Evaluation Managers

- Keep in mind that there is no perfect design, the evaluator will need to choose the design based upon trade-offs of time, cost and practicality.
- Assist the evaluator with this decision, by using information about budget, timelines, and importance of the intervention or policy to be evaluated.

Application Exercise 6-1

Selecting an Evaluation Design

Scenario:

You have been asked to measure the impact of building a community health clinic to teach parents how to treat common family illnesses, and how to identify something that might be more serious. The goals are to increase the number of parents with basic understanding of preventative healthcare, first aid, and early treatment strategies, and to reduce the number of children and elderly whose illnesses become serious.

1. What is the program?

2. What are the desired outcomes?

3. How would you write an cause-effect question for this evaluation?

4. What design might you use?

5. Why? What are the strengths and limits of this design? Why is this better than other possible designs?

Application Exercise 6-2

Selecting an Evaluation Design and Data Collection Strategy

Scenario:

You have been asked to create an evaluation design for a six-month study to assess the effectiveness of a preventative health campaign in your country. You have a moderate budget that will allow some assessment of outcomes, and you have a team of six research assistants to help you with the details. This campaign is to consist of two-day seminars conducted by health professionals in communities throughout your country. The purpose of your evaluation is to determine whether the campaign resulted in improved health practices by citizens.

Is your primary evaluation question a descriptive, a normative, or an cause-effect question? Explain.

Should your data collection strategy be more structured, more open-ended, or a combination of both? Why?

How would you identify the most important outcomes to measure, and how would you measure them?

What evaluation design elements would you use (e.g., inclusion of a comparison group, controlling for other variables, causal tracing strategies, etc)? What are strengths and weaknesses associated with your design?

Further Reading and Resources:

Campbell, D.T. and Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. In N. L. Cage (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally

Cook, T.D. and Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Davidson, E.J. (2000). Ascertaining causality in theory-based evaluation. *New Directions for Evaluation,* No. 87:17-26.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Patton, Michael Q. The debate about randomized controls in evaluation: The gold standard question. PowerPoint presentation to IPEDET, June 2005.

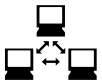
Stake, R.E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage Publications.

Stufflebeam, D. L., Mdaus, G. F., and Kellaghan, T., editors (2000). Evaluation Models: Viewpoints on educational and human services evaluation. Boston: Kluwer.

Wadsworth, Y. (1997). *Everyday evaluation on the run*. St. Leonards, NSW, Australia: Allen and Unwin.

Yin, R.K. (1984). *Case study research*. Thousand Oaks, CA: Sage Publications.

Websites:



The Campbell Collaboration: http://www.campbellcollaboration.org/

Kisker, E.E., & Brown, R.S. (1997). Nonexperimental Designs and Program Evaluation. *Children and Youth Services Review 19,* no. 7 (1997): 541-66. Online:

http://www.aei.org/publications/pubID.17770/pub_detail.asp

NIOSH/A Model for Research on Training Effectiveness (TIER). Online: http://www.cdc.gov/niosh/99-142.html

Trochim, W. Designing designs for research. Online:

http://www.socialresearchmethods.net/kb/desdes.htm

Answers to Quiz Yourself

1. d
2. d
3. d
4. a
5. b
6. a
7. d
8. c
9. d
10.
    - There is no perfect design.
    - Each design has strengths and weaknesses.
    - There are always trade-offs and potential weaknesses of each design.
    - Acknowledge trade-offs and potential weaknesses in your design report.
    - Provide some assessment of the likely impact of your design choice on your results and conclusions.
    - Design choices drive subsequent methods of data collection and analysis.

To continue on to this
Module's Presentation Slides click here

To continue on to the
Next Module click here

To return to the
Table of Contents click here