Evaluation in Organizations

A Systematic Approach to Enhancing Learning, Performance, and Change

2nd Edition

Darlene Russ-Eft
Hallie Preskill

BASIC
BOOKS

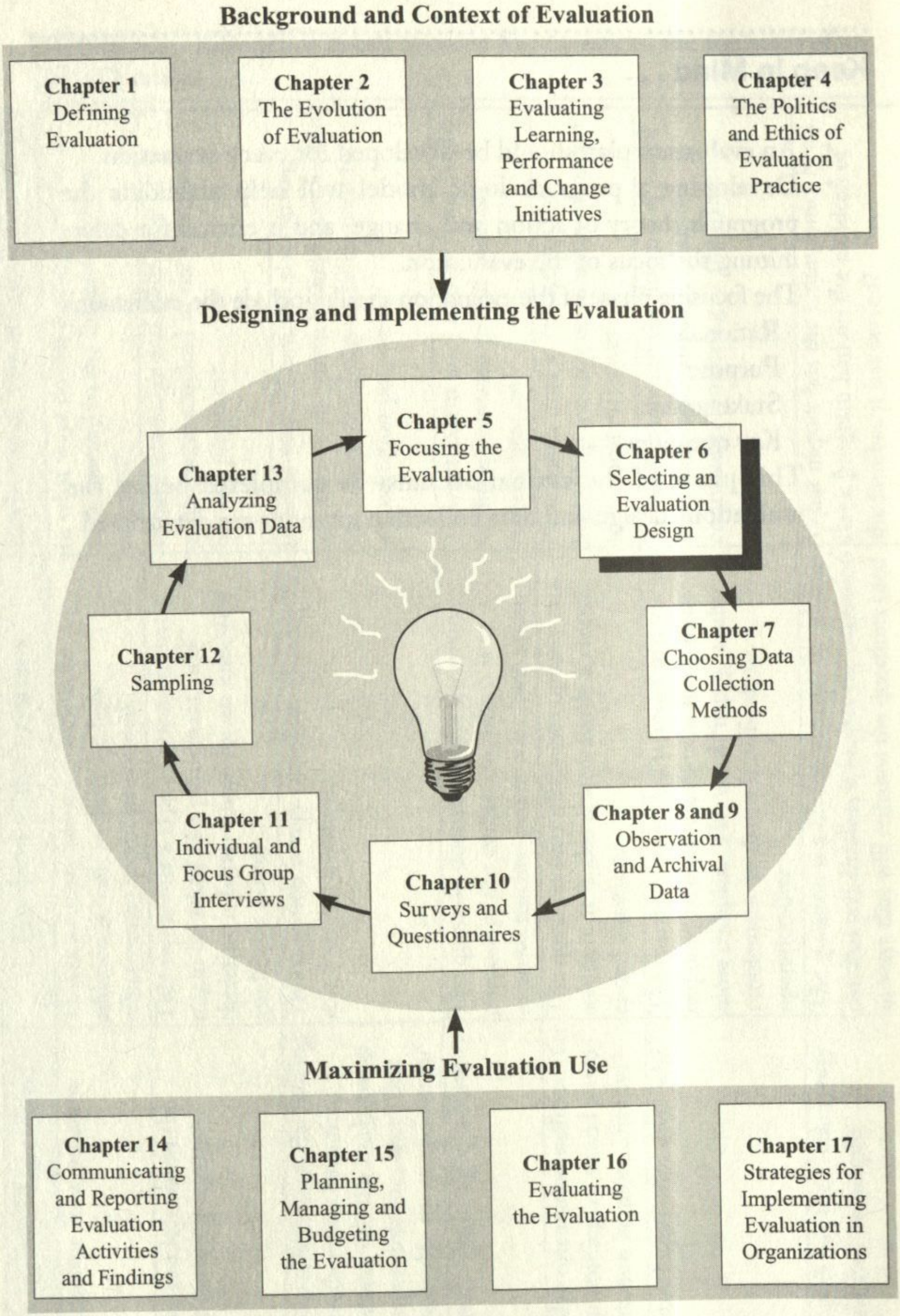
A Member of the Perseus Books Group
New York

Published by Basic Books,
A Member of the Perseus Books Group

Designed by Linda Mark

A CIP catalog record for this book is available from the Library of Congress.
ISBN: 978-0-465-01866-6

10 9



Background and Context of Evaluation

Chapter 1
Defining Evaluation

Chapter 2
The Evolution of Evaluation

Chapter 3
Evaluating Learning, Performance and Change Initiatives

Chapter 4
The Politics and Ethics of Evaluation Practice

Designing and Implementing the Evaluation

Chapter 5
Focusing the Evaluation

Chapter 6
Selecting an Evaluation Design

Chapter 7
Choosing Data Collection Methods

Chapter 8 and 9
Observation and Archival Data

Chapter 10
Surveys and Questionnaires

Chapter 11
Individual and Focus Group Interviews

Chapter 12
Sampling

Chapter 13
Analyzing Evaluation Data

Maximizing Evaluation Use

Chapter 14
Communicating and Reporting Evaluation Activities and Findings

Chapter 15
Planning, Managing and Budgeting the Evaluation

Chapter 16
Evaluating the Evaluation

Chapter 17
Strategies for Implementing Evaluation in Organizations

Chapter 6

Selecting an Evaluation Design

- Basic Design Issues
- Commonunly Used Evaluation Designs
- One-shot Design

Vignette 3: Testing to No Avail at Good Foods, Inc.

Tim Jenkins, head of the company's human resource development department, wanted to conduct an evaluation of a newly designed sales training program. Since the program was supposed to increase the participants' knowledge of specific sales techniques, he decided to administer a test immediately after the last training session to see how much they had learned. He assumed that increased knowledge would result in increased sales revenue. The test results showed consistently high scores. But when Tim showed the results to the vice president of sales, the vice president expressed some skepticism about the usefulness of results and said, "A lot of these folks already knew this stuff."

Tim might have avoided this criticism by using a better evaluation design, specifically one that controlled for prior knowledge. Instead of simply using a knowledge test following training, he might have considered administering a similar test before the training program. This would have provided valuable information on what the participants actually knew before beginning the sales training program. Alternatively, he might have randomly selected two groups of salespeople, with one group going through the training first. Then he could have administered the knowledge test to both groups at the same time, when the first group had finished training but the second group had not yet started. This would have provided information on the differences in knowledge between those who had completed the training and those who had not yet started.

A variety of evaluation designs currently exist, and one or more of these designs may be appropriate for evaluating a learning, performance, or change initiative. This chapter will provide an overview of some of the most commonly used types of evaluation designs and highlight their strengths and weaknesses. Before reviewing these designs, however, we discuss some basic issues relating to evaluation design. The issues discussed in this chapter are relevant for both evaluators and researchers.

Basic Design Issues

Before jumping into a discussion of different designs and related issues, we want to identify two different views of evaluation and research. One perspective has been called scientific, quantitative, empirical, positivistic, and post-positivistic. In this view, the evaluator or researcher stands independent and apart from what is being studied. Furthermore, she or he must adopt a neutral stance and control for any biases or preconceived ideas. The purpose of these kinds of studies is to identify cause-and-effect relationships and to discover laws that govern these relationships.

The second view has been called qualitative, ethnographic, and naturalistic. Those who adopt this view believe that evaluators cannot separate themselves from what is being studied—they cannot help but bring their subjectivity and values to the evaluation. The purpose or objective for qualitatively oriented evaluations is to understand, as much as possible, the lived experiences of those being studied.

We have found that most evaluators tend to adopt one view or the other, typically because of their background, education, or training. As a result, these factors often influence the evaluation's key questions, design, and choice of data collection methods. To ensure that both worldviews and preferences are represented, or that one perspective is not embraced to the exclusion of the other, it is a good idea to conduct the evaluation using a team-based approach. Such a team would consist of evaluators possessing quantitative and qualitative orientations and expertise. This would then facilitate the collection of both kinds of data (often referred to as "mixed methods").

Design Validity

The usefulness of evaluation findings is often based on the accuracy and credibility of the results. In this section, we discuss the concepts of internal and external validity for quantitative and qualitative data.

Internal Validity. The internal validity of an evaluation effort refers to the extent to which it correctly answers the questions it claims to answer about what is being evaluated. Do the data accurately represent how people think, feel, know, or act about the evaluand? All data collection and analysis efforts are carried out in the context of a set of assumptions (sometimes considered a model) about the program being observed. If those assumptions are wrong, then the findings of the evaluation are meaningless. For example, we might assume that a program that trained people to run a mile faster would improve their performance as a team leader. The evaluation might then focus on whether the training led to faster running times, simply assuming that this would result in improved team leadership. If those assumptions are correct, then the evaluation is internally valid, and the findings from the evaluation are meaningful. Note, however, that the example is one where the assumptions are wrong and therefore the findings are meaningless.

In evaluations, whether quantitative or qualitative approaches are used, another major threat to internal validity is that unmeasured processes (also called confounding factors or variables) might account for the results that were obtained. When learning, performance, and change initiatives are evaluated, the results may owe to confounding variables such as previous knowledge or history with the organization or the job. (This was the criticism leveled by the vice president of sales in our beginning vignette.) The design of the evaluation can help account for such factors or variables. Confounding variables might include the following (words in italics are the technical terms used to describe a particular threat to validity; adapted from Campbell and Stanley 1963):

- Specific and unexpected events affecting the variables of interest (such as a major downsizing initiative)—*history.*
- Passage of time that leads to changes in attitudes or behavior (such as time on the job resulting in the acquisition of certain skills)—*maturation.*
- Effects of one data collection effort on later data collection (such as the experience of taking a test affecting performance on a later test)—*testing.*
- Changes in the data collection instruments or the observers, which may affect the manner in which measurements are taken (such as different observers at different times or changes in item wording on a postsurvey)—*instrumentation.*
- Attrition from the sample (such as job or organizational changes)—*mortality.*

A second type of threat to internal validity that can affect both quantitative and qualitative data involves using data collection methods that do not accurately measure the underlying dimensions of concern. Let's say a person being hired to fill a clerical position is given an employment test that measures driving ability. In this case, the test scores are irrelevant for making judgments about the quality of the job candidate. Or let's say you are conducting interviews that focus on issues of workplace discrimination. The interview questions may fail to obtain needed information because of respondents' desire to be politically correct, or a fear of repercussions if they were to provide their real opinions. In this situation, more attention to the types of questions being asked may reduce problems with internal validity.

In some evaluations, comparisons are made between a group that receives the learning, performance, or change intervention (sometimes called the experimental or treatment group) and a group that does not receive the intervention. In comparison designs, another threat to internal validity is the equivalence between the two groups. In most learning, performance, and change evaluations, treatment and control groups have not been randomly selected and assigned from a predetermined population. Such nonrandom assignment can lead to systematic distortion of the results, particularly if those selected for the intervention are the "best" or the "worst" performers, the "most motivated" or the "least motivated" employees. In addition to problems related to a lack of initial random assignment, the two groups may become nonequivalent over time because more people drop out of one of the two groups. The primary methods for ensuring internal validity are those used to conduct most research studies: (1) random assignment to treatment and control conditions, (2) identification and measurement of confounding factors, (3) control over confounding factors, and (4) the use of multiple methods, such as rating forms and interviews, to obtain converging evidence in support of a particular finding (also known as *triangulation*). If an evaluation lacks internal validity, the evaluation will lose credibility in the face of any serious criticism. In such cases, decision makers will turn to other information sources to make important decisions, and the time, effort, and cost of the evaluation will be viewed as a worthless expense.

There are, however, very real costs for increasing an evaluation's internal validity. One example is the increased direct costs for measuring confounding effects. This includes hiring one or more experts to consider all these confounding variables beforehand, the cost of developing and piloting methods for measuring confounding variables, and the costs of collecting and analyzing the additional data.

Therefore the evaluator must decide whether an evaluation effort needs to incur the costs for increasing internal validity. A high level of internal validity is necessary if the evaluation results are to be used for important decision-making purposes. On the other hand, internal validity need not be as great for exploratory evaluations. You may be undertaking an evaluation to pilot some data collection methods or to explore some tentative ideas concerning a particular evaluation issue. In such cases, you may not need high levels of internal validity, because you plan to undertake further evaluation to test conclusions that were tentatively reached from an exploratory study. This does not mean that exploratory evaluations can be conducted without pilot tests or concerns for data accuracy. A certain level of internal validity is always needed to justify the collection and analysis of data.

At the onset of the evaluation, you should make a rough determination as to the level of validity needed. If it is unattainable, a decision should be made as to whether to conduct the evaluation. For example, imagine that you have been asked to conduct an evaluation of an organization's performance management system at the same time as the organization is being merged with another company. In deciding whether to conduct the evaluation, you should consider not only the extent to which employees will agree to be interviewed or surveyed, but you must also judge how honest their responses will be during this unsettled time. You may conclude that given employees' mistrust of management, the resulting data may not be valid and consequently choose not to pursue the evaluation until things settle down.

External Validity. This form of validity concerns the extent to which the results can be generalized to other situations. For example, can the results from one site be generalized to the entire organization or to other organizations? Generally, external validity is determined by *sample selection*, whereas internal validity is determined by *sample assignment*. Thus external validity is increased to the extent that the sample selection reflects the population and the results can be generalized. In contrast, internal validity is increased through random assignment of the selected sample to one or more groups.

Such random assignment may not in fact be the method by which participants are assigned to learning, performance, and change interventions. In many situations, people in the organization may be assigned to particular learning or training opportunities because of their location or because they are viewed as "rising stars." So, if you used random assignment in the evaluation, you would increase the internal validity and reduce external validity and generalizability. In other words, you would have obtained high-quality results (internal validity), but these findings would generalize only to situations where people were assigned randomly (external validity).

Factors that can affect a study's generalizability include:

- Reactions to the pretest that makes the results unrepresentative of the general population. In this case, people might be sensitized to certain issues because of the pretest, and such sensitivity would not extend to the larger population.
- Reactions to "experimental" conditions, which again make the results unrepresentative of those outside such conditions. As with the previous factor, such sensitivity or reactions may accompany the specific training setting but would not extend to other learning and performance settings.

- Reactions to multiple treatments, which makes the results unrepresentative of those not receiving the same set of treatments. Many learning, performance, and change initiatives involve more than a single event in isolation. In such cases, the results may not extend to other kinds of interventions.

The value of external validity is the ability to generalize the results to a larger population. Such generalizability, then, depends on drawing a representative sample of the population. (See Chapter 12 on sampling.) Another approach to maximizing external validity is to repeat the evaluation (replication).

As with internal validity, the expense for achieving external validity comes in the form of direct costs. Obtaining a broad and representative sample involves several technical procedures and may require the cost of involving an expert in sampling procedures. Replicating an evaluation would require the costs for repeating the data collection and data analysis processes.

Multicultural Validity. Kirkhart (1995) introduced the construct of multicultural validity as "the accuracy or trustworthiness of understandings and judgments, actions and consequences, across dimensions of cultural diversity" (p. 1). Such dimensions can include demographic aspects (i.e., race and ethnicity, gender, language, age, religion, sexual orientation), sociopolitical factors (i.e., power, economic status, class), and context-specific features. Evaluators must recognize that each program or process being evaluated is situated within a particular culture and must consider issues of culture throughout an evaluation's design and implementation.

An important aspect of multicultural validity involves the inclusion of marginalized groups. According to Weiss (1998), such groups "are apt to have different interests and concerns . . . and addressing their questions would broaden the scope of the study" (p. 30). Mertens (1999, p. 6) also suggests that "inclusive evaluation has the potential to contribute to an enhanced ability to assert truth, objectivity, credibility, validity, and rigor in the sense that ignored or misrepresented views are included." Similarly, Hopson (2001) states:

> Especially in traditionally underserved communities, we would do well to reconsider how these communities' knowledge ought to be viewed in the

context of evaluation thinking and design, and how certain methodologies and tools would work best in an evaluation. (p. 379)

The concept and practice of ensuring multicultural validity is of particular concern as one undertakes an evaluation in global organizations, with indigenous peoples, and in multicultural communities. The evaluator must consider what designs, approaches, and data collection methods are most credible, appropriate, and meaningful within the evaluation's cultural context.

Validity and Qualitative Data. What we have just discussed is the concept of validity primarily in relation to quantitative methods and designs. Again, validity relates to the degree of confidence we have that the data represent our participants' truth or reality. What constitutes "truth," however, is defined in different terms when we use qualitative approaches and methods.

Guba and Lincoln (1981) propose that evaluators and researchers using qualitative designs should still be held accountable for how well they depict the reality of respondents. They recommend that evaluators develop and use "tests of rigor" for establishing such validity. As examples, they provide four criteria of rigor for judging the trustworthiness (validity) of qualitative data.

1. *Truth Value: How can one establish the confidence in the "truth" of the findings of a particular inquiry for the subjects with which—and the context within which—the inquiry was carried out?* Whereas the "scientific" or "positivistic" paradigm asserts that there is one reality and that information is valid when all relevant variables can be controlled and manipulated, a naturalistic or qualitative paradigm assumes that there are multiple realities that exist in the minds of individuals. Thus, when using qualitative methods, the evaluator seeks to establish the credibility of individuals' responses. The study must be believable by those who provide the data and by those who will use its findings. Providing a detailed, comprehensive depiction of the multiple perspectives that exist can enhance the data's credibility. For example, trainee satisfaction ratings from a survey, along with interviews of trainees' managers and the training instructors, would provide a more holistic and credible picture of the training experience.

2. *Applicability: How can one determine the degree to which the findings of a particular inquiry may have applicability in other contexts or with other subjects?* How transferable or applicable the findings are to another setting is called generalizability in the scientific paradigm. The goal for qualitative methods is to provide a richly detailed description; such a description can help the reader relate certain findings to his or her own experience. We often think of these as "lessons learned." For example, as a stakeholder reads an evaluation report, he realizes that something very similar has occurred in his own organization, and sees where some of the findings can be used. Although the entire set of findings may not be applicable to his context, some issues identified or lessons learned have applicability in other contexts.

3. *Consistency: How can one determine whether the findings of an inquiry would be consistently repeated if the inquiry were replicated with the same (or similar) subjects in the same (or a similar) context?* In the scientific paradigm, the notion of consistency is called reliability, where a study or instrument's consistency, predictability, or stability is measured. Since reliability is necessary for validity, it is critical that data of any kind be reliable. Instead of considering data unreliable if it is inconsistent, evaluators using qualitative methods look for reasons that the data appear unstable (inconsistent). For example, an interviewee might give an opinion one day, and when asked that same question again the following week, might say something slightly different. What would be important to understand and capture are the reasons for the change in perception. Such inconsistencies may stem from respondent error, an increase in available information, or changes in the situation. An audit trail that includes collecting documents and interview notes and a daily journal of how things are going can help to uncover some of the reasons for such inconsistencies.

4. *Neutrality: How can one establish the degree to which the findings of an inquiry are a function solely of the subjects and conditions of the inquiry and not of the biases, motives, interests, and perspectives of the inquirer?* Objectivity is often viewed as the goal of most evaluation and research studies. Evaluators and researchers who use qualitative methods don't necessarily believe, however, that true objectivity can ever be fully achieved. They believe that it is impossible to completely separate the evaluator from the method. Instead of trying to ensure that the data are free of the evaluator's biases,

the goal is to determine the extent to which the data provide confirming evidence. "This means that data (constructions, assertions, facts, and so on) can be tracked to their sources, and that the logic used to assemble the interpretations into structurally coherent and corroborating wholes is both explicit and implicit" (Guba and Lincoln 1989, p. 243). Establishing confirmability, like consistency, often takes the form of auditing.

To help understand the relationship between these tests of rigor and the scientific terms for validity, Guba and Lincoln (1981, p. 104) provide the following chart:

Aspect	*Scientific Term*	*Naturalistic Term*
Truth value	Internal validity	Credibility
Applicability	External validity	Fittingness generalizability
Consistency	Reliability	Auditability
Neutrality	Objectivity	Confirmability

Techniques for Establishing the Validity of Qualitative Data

To ensure that you do everything possible to enhance the validity of evaluation data, consider implementing the following techniques:

- Carefully check the accuracy in data recording and coding—ensure that all of the data are ready and available for analysis.
- Make repeated and persistent observations—build trust and rapport with participants so they are more likely to provide valid information. The longer one is on site, the more trust can be built, and the more complete the observations will be.
- Member checking is a means for testing categories, interpretations, or conclusions through continuous checking of data with stakeholders or other participants after various data collection activities, or when a draft of a final report has been written.
- Triangulation is a means for contrasting and comparing information drawn from different sources, methods, and theories. The term triangulation comes from land surveying and navigation where multiple reference points are used to locate an object's exact position. "Knowing a single landmark only locates you

somewhere along a line in a direction from the landmark, whereas with two landmarks you can take bearings in two directions and locate yourself at their intersection" (Patton 1990, p. 187). Therefore the more information you have from as many sources as possible, which have been subject to various interpretations, the greater the likelihood the data are trustworthy. Denzin (1978) describes four kinds of triangulation that are commonly used today.

1. *Data triangulation*—collecting data from a variety of sources. For example, in evaluating the transfer of learning from a three-hour workshop, the evaluator collects information from the training participants, their managers, and their peers. In this case, three different sources have been queried.
2. *Methodological triangulation*—using more than one method to collect data. For example, we may conduct individual interviews with 20 percent of a department's employees and survey the remaining 80 percent. By using two methods, the weaknesses of one may be compensated by the other.
3. *Investigator triangulation*—involving two or more evaluators in the inquiry. The maxim "Two heads are better than one" applies to this kind of triangulation. When two or more evaluators can plan, implement, debrief, and interpret the findings, the data will likely be richer and more trustworthy.
4. *Theory triangulation*—using different theoretical perspectives to interpret the same data. By applying different theories to make sense of the data, it is possible to see how different assumptions and beliefs influence one's interpretations. By making these explicit, stakeholders can see how their assumptions might influence various actions taken because of the findings.

- Peer debriefing involves engaging in formal or informal discussions with a peer about what was seen, heard, experienced, and interpreted. This helps explore alternative explanations and emerging themes and patterns.
- Audit trail allows an auditor to determine the trustworthiness of the study. The evaluator must retain and make available interview guides, field notes, documents, peer debriefing notes, journals,

and any other documents that have been used or collected. The auditor then may review these documents as part of the audit.

- Pilot testing means trying out each of the data collection instruments with a sample of the population (or one similar to it). This enables the evaluator to determine if the questions are likely to elicit the kinds and quality of information being sought. Revisions can be made before implementing the instruments with the total sample or population.
- Rival explanations method refers to a process for looking for multiple ways of organizing the data that might lead to different findings.
- Negative cases involves considering cases that do not fit a pattern. These cases may lend insights into emerging issues or new perspectives on a recurring problem.

Commonly Used Evaluation Designs

In the following section we describe various types of evaluation designs. After explaining each one, we provide an example and then list the design's strengths and weaknesses. Further details on these and other designs can be found in Campbell and Stanley (1963), Cook and Campbell (1979), and Russ-Eft and Hoover (2005).

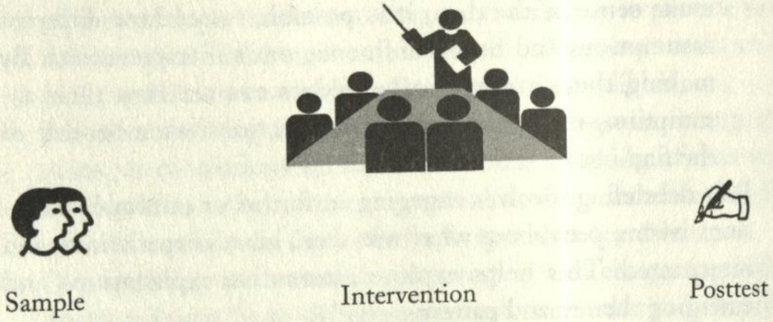


ILLUSTRATION 6.1 One-Shot Design

One-Shot Design

The term "one-shot" refers to the fact that the measurement takes place at one time only. This can be depicted as shown in Illustration 6.1.

The one-shot design is commonly used in evaluating learning, performance, and change interventions, whereby after the event, a postcourse or postevent survey is administered to participants. This design assumes that the respondents are providing reactions to the intervention, and only to the intervention; they are not reacting to the latest announcements about organizational change or other aspects of their environment.

Another example of the one-shot study involves measuring learning and performance outcomes following training on a specific procedure or piece of equipment, such as was described in the vignette at the beginning of the chapter. In that case, Tim decided to measure the effects of sales training by using a test at the end of the workshop. Thus if the content of the test was not known before the training, it may be reasonable to assume that the correct description of how to implement those steps as recorded on the test resulted from the training. Furthermore, if the trainees do indeed remember the relevant course material for the test, it is often assumed that they can successfully apply the material to the job.

A major advantage of this design is that it is simple and cost-effective. A limited amount of data is collected at one time only. In many cases, these data can be gathered as part of the learning, performance, and change initiative.

One-shot designs make several assumptions, however. They assume that the measurement is related in some way to the intervention rather than to some other factor. That assumption may or may not be accurate. For example, negative reactions to a training program may result from a recent organizational communication regarding downsizing and may have nothing to do with the training program's quality. As another example, the correct use of some procedure or equipment may be the result of previous knowledge and experience and not the training. However, if the procedure or equipment had not been previously used, it may be assumed that correct usage resulted from the intervention. A final problem with the one-shot design involves the low level of external validity. Since this takes place with only one group, it is not clear whether the results can be generalized to other populations. Such generalization may be addressed if participants were randomly selected for the intervention.

Advantages of the one-shot design:

- Is simple and cost-effective to conduct
- Reduces the costs and time for data collection and analysis
- Produces data that can be analyzed in a timely and cost-effective way
- Can gather data as part of the learning, performance, or change event
- Provides needed information when using an ideal comparison

Disadvantages of the one-shot design:

- Does not control for the effects of other factors, such as organizational issues or prior knowledge
- Assumes that positive reactions and knowledge tests lead to behavior changes
- Findings do not necessarily generalize to other populations

Retrospective Pretest Design

In this variation of the one-shot case study design, data are collected from participants following the learning, performance, or change intervention; however, the participants report retrospectively on their attitudes, knowledge, or skills, as well as report on their current attitudes, knowledge, or skills. Illustration 6.2 depicts this design. As a result, the evaluator can compare these retrospective preassessments with their postassessments.

The following is an example of the type of item that Tim might use in a survey of trainees when evaluating behavior change as a result of the sales training (See Table 6.1).

Indeed, one could use the simple type of form above or a 5-point or 7-point rating scale.

TABLE 6.1 Example Retrospective Pretest Study

For each of the following sales skills, please indicate (1) your skill level prior to training and (2) your current skill level.

Sales Skills	Skill Level Prior to Training			Current Skill Level		
	Poor	Average	Excellent	Poor	Average	Excellent
Making cold calls						
Identifying customer needs						
Presenting solution						
Preparing proposal						
Closing the sale						

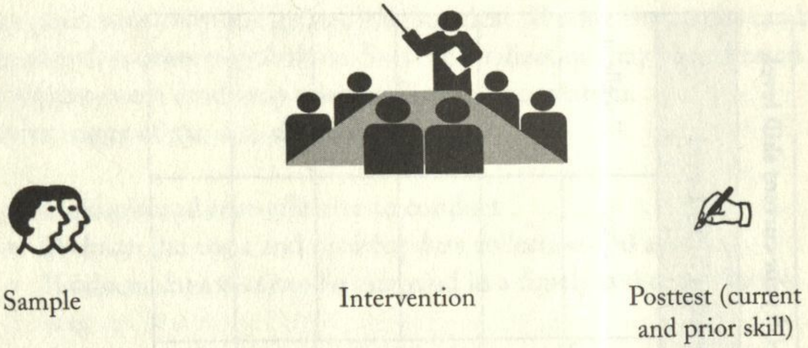


ILLUSTRATION 6.2 Retrospective Pretest Design

This design depends on the accuracy of participants' recall, as well as their willingness to provide "truthful" data. For example, in communication skills training, trainees may or may not be aware of their prior skill level until after the training. In these cases, the retrospective design may provide a more accurate picture of pretraining skills than data gathering before training.

Another example of this design would be the administration of a survey measuring attitudes toward customers following participation in a new flextime work program. The survey would ask the respondents about their attitudes before the program as well as their current attitudes. In the case of the sales training vignette at the beginning of this chapter, rather than asking about specific skills, Tim could ask participants about their attitudes and practices prior to the program and now, six months after the program's initial implementation. Another variation of this design is to ask participants to describe what impact it has had on their intention to remain in the organization.

Like the one-shot design, the appeal of the retrospective design is its simplicity and ease of data collection. In addition, you can obtain a comparison between posttest data and the retrospective predata. Furthermore, the posttest data are not contaminated by the experience of pretesting.

One drawback of the retrospective design is that it does not include a control group of people who did not participate in the intervention. As a result, the evaluator cannot rule out the possibility that the results were due to history with the organization or with the job (as is true with the one-shot design). Another problem, particularly with the retrospective reporting, involves possible distortions in the retrospective reports. Such distortions may result from memory problems or from the respondents' current attitudes and beliefs. In some cases, actual preassessments collected from participants, their managers, their peers, and their subordinates often produce inflated scores as compared with retrospective preassessments. In such cases, the actual pretest scores compared with posttest scores may show little or no difference; in contrast, the retrospective pretest scores compared with posttest scores may show significant differences. Again, the use of or the lack of random assignment to the intervention may increase or decrease the external validity of this design. Taylor, Russ-Eft, and Taylor (2009) identify some biases that can result in greater prepost differences with the retrospective pretest than with an actual pretest. They suggest that items assessing skills not included as part of the intervention or problem be used in the assessment as a type of "control." Presumably there should be no differences in the control items, and if there are, these differences can show the level of bias that may be occurring.

Advantages of the retrospective pretest design:

- Is simple and cost-effective
- Reduces the costs and time for data collection and analysis
- Gathers data as part of the learning, performance, or change event
- Compares posttraining data with retrospective predata
- Avoids attrition from the sample being tested or measured
- Decreases the likelihood of testing effects

Disadvantages of the retrospective pretest design:

- Cannot rule out the possibility that results were owing to history with the organization or the job (as is true with the one-shot designs)
- Possible distortions in retrospective reports because of memory or current attitudes and beliefs

One-Group Pretest-Posttest Design

This design involves data collection before the learning, performance, or change intervention as well as following it. This can be depicted in Illustration 6.3.

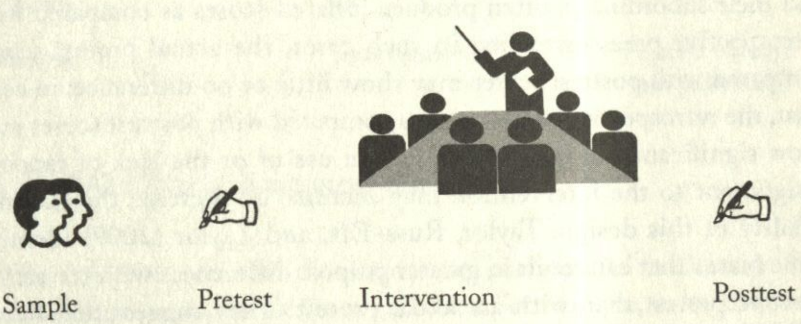


ILLUSTRATION 6.3 One-Group Pretest-Posttest Design

Using the sales training example again, Tim could test trainees on sales protocol procedures before and after the sales training course. This would provide information regarding how much trainees had learned.

A typical scenario for a posttest is to administer it immediately following the learning intervention, perhaps even on the last day of the event. Depending on what is being measured, such timing may or may not be appropriate. For example, if you want to measure whether trainees display their newly developed skills on the job, you would want to allow trainees some time to use and become comfortable with these skills. Or you may measure participants' reactions both immediately following the training and some time later to see if their perceptions changed with the passage of time.

If you decide to collect data over a period of time, it is important to consider how often data need to be collected (see the time-series design later in this chapter). For example, you might decide to conduct certain observations with greater frequency, such as once a month, rather than less frequently, because you believe that there will likely be significant variation of individual or group performance within the longer period. Annual reports to Congress or to stockholders in a corporation provide examples of a yearly evaluation of certain government programs or of certain aspects of a corporation's operations. On the other hand, economic measures relating to our society, such as cost-of-living indicators, fluctuate seasonally, so these measurements occur several times a year.

Many evaluations of learning, performance, and change initiatives tend to be undertaken on a one-time basis, in which there is no plan for repeated data collection. This may or may not be appropriate. If the initiative is long-term and continues over many years, you may want to consider conducting some evaluation activities on a periodic basis. In such cases, you might consider planning the initial study in such a way that the data provide a baseline for long-term studies.

As with the previous designs, this design is relatively simple and cost-effective. Indeed, the participants could be asked to complete an instrument that focuses on attitudes, opinions, knowledge, or skill level at the beginning and at the end of the learning, performance, or change intervention. Since data would be collected as part of the intervention, the evaluator would reduce the costs of data collection and the possibility of attrition from the sample.

However, there are several possible weaknesses with this design. Without a control group of people who have not participated in training, similar problems as those mentioned with the one-shot and retrospective designs can arise. These include the possibility that the results appear because of previous history with the organization and the job (history and maturation). In addition, the pretest itself may cause changes to occur irrespective of the training (testing). For example, a knowledge or skills pretest may actually contribute to the improved knowledge or skills of the people taking the posttest. If so, the posttest measurement not only reflects the effects of the intervention, but it reflects the effects of the pretest *plus* the intervention. Another potential problem involves that of increased effort and resources for follow-up and possible attrition or loss of people from the sample (mortality). With designs that require repeated data collection, you may find respondents disappearing from their positions and from the organization, resulting in a smaller than expected sample (mortality). This becomes particularly problematic when the posttest must take place at some time following the learning and performance intervention. Finally, as with the previous designs, random assignment to the learning, performance, or change intervention can increase or decrease the generalizability of the findings, depending on the organizational circumstances.

Advantages of the one-group pretest-posttest design:

- Can be simple to conduct
- Can be cost-effective
- Reduces the costs and time for data collection and analysis
- Gathers data as part of the learning, performance, or change event
- Measures actual attitudes, knowledge, and skills prior to the intervention
- Allows for actual comparison of pretest and posttest data

Disadvantages of the one-group pretest-posttest design:

- Cannot rule out the possibility that results were due to history and maturation with the organization or the job (as is true with the one-shot designs)
- Cannot rule out instrumentation effects (because of changes in the instruments or observers)
- Pretest itself may cause changes to occur irrespective of the training
- Is vulnerable to the loss of people from the sample because of job changes
- May require added data collection costs for conducting the posttest

Posttest-Only Control Group Design

Although researchers and evaluators usually emphasize the importance of a pretest, you may want to consider avoiding the use of the pretest to eliminate the effects of the pretest on the posttest results. One can do this by using the posttest-only control group design. See Illustration 6.4.

In such a design, two groups may be randomly selected, with one group experiencing the learning, performance, or change intervention, and the other receiving no intervention. The two groups are then given a posttest at the same time following the intervention. So, going back to our original vignette, Tim could randomly assign salespeople to one of two groups. One group would receive training, and the other group

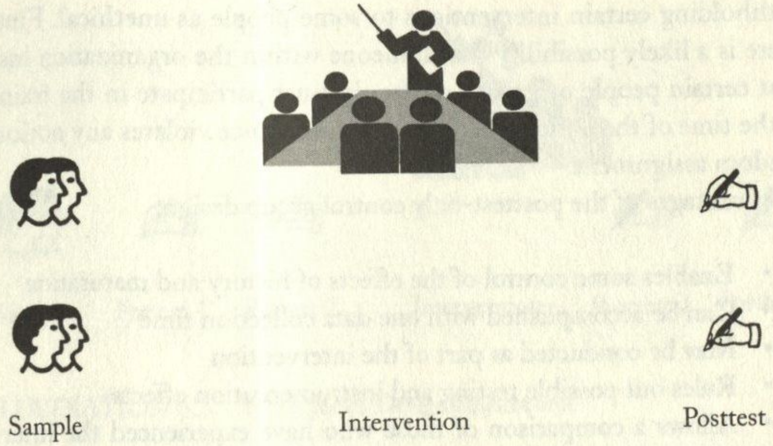


ILLUSTRATION 6.4. Posttest-Only Control Group Design

would not receive training (or would receive training at some later time). Following training, the two groups would complete the same posttests.

As previously mentioned, this design avoids the problems associated with using the pretest. Furthermore, random assignment to the two groups provides assurance that no systematic bias exists between the two groups. Concerns over such systematic bias usually lead evaluators to recommend the use of a pretest to determine the similarity of the two groups. Random assignment can, however, overcome this problem, because you can assume that the groups are the same. In addition, data collection is made more simple and easy since it takes place at one time only.

The major limitation, certainly within organizations, is the lack of feasibility of random assignment to various treatment conditions. This problem is exacerbated when you attempt to assign certain people to a "control" condition where they do not receive the intervention. (One effective alternative in a larger organization is to suggest that the "control" group receive the intervention later on, after the posttest.) For organizations with the imperative of completing everything *now*, the notion of keeping some people from needed developmental experiences seems inappropriate. In addition, some organizations may consider withholding certain interventions to some people as unethical. Finally, there is a likely possibility that someone within the organization insists that certain people or groups of people must participate in the training at the time of the evaluation. If so, such insistence violates any notion of random assignment.

Advantages of the posttest-only control group design:

- Enables some control of the effects of history and maturation
- Can be accomplished with one data collection time
- May be conducted as part of the intervention
- Rules out possible testing and instrumentation effects
- Allows a comparison of those who have experienced the intervention with those who have not

Disadvantages of the posttest-only control group design:

- Assumes random assignment to treatment and control groups, which may prove impractical in an organizational setting
- Raises potential ethical problems when withholding the intervention from a certain group
- May require added data collection costs for conducting the posttest
- Does not rule out the possibility that everyone knew this information or had these skills prior to the intervention

Time Series Design

So far, we have only looked at two separate data collection times—a pretest and a posttest. Another possible design involves repeated data collection before and following the learning, performance, or change intervention. See Illustration 6.5.

Using this design, one would graph the results obtained at each time. If charting the results, such as sales revenue, showed some dramatic change only from the time immediately before to immediately after training, one could assume that the training had some impact on the results. Conducting several pretests establishes a stable baseline to use for comparing

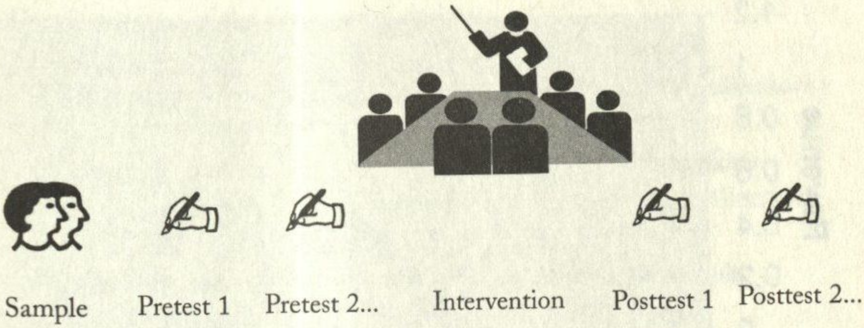


ILLUSTRATION 6.5 Time Series Design Illustration

posttraining results. Let us return to the vignette at the beginning of this chapter. Let's say Tim decides to chart the revenue of salespeople over time. If training took place in early April and the graph of sales revenue looked something like that shown in Illustration 6.6, Tim would have reasonably strong evidence that the sales training program had some effect.

A strength of this design is the fact that you have shown that time in the organization or on the job by itself did not result in changes in attitudes or behavior. For example, let us assume that you measured performance each month for six months showing a gradual improvement over time. Then you show a large and significant improvement in performance following an intervention. Time or history alone would probably not account for such a large improvement unless accompanied by the intervention (or some specific organizational change).

There are, however, several problems with this design. First, as mentioned above, you cannot easily isolate various organizational influences that could affect one's performance separate from the learning, performance, and change intervention. In addition, you would have to undertake repeated data collection with the participants. Such repeated measurements lead to additional costs for the evaluation. Also, the repeated measurement could result in changes in attitudes or behavior simply because of the measurement itself. Since this design takes place over time, there may be attrition from the sample because of people being reassigned or

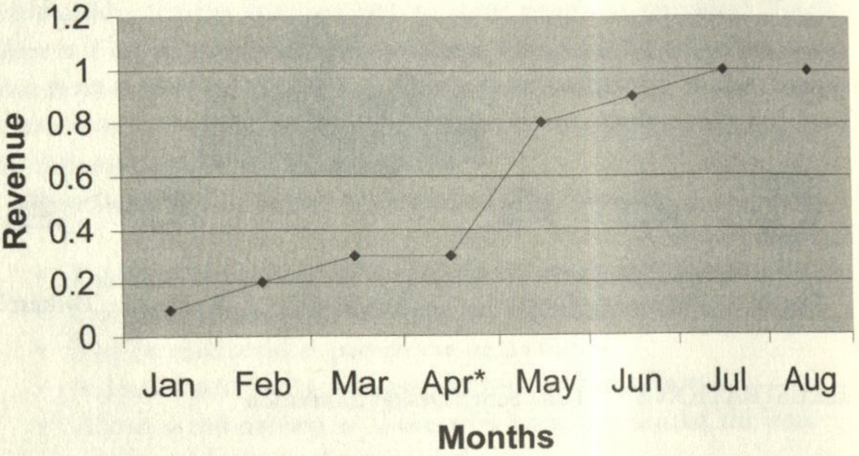


ILLUSTRATION 6.6 Sales Revenue over Time

leaving the job and the organization. In addition, whenever someone responds multiple times to the same data collection instrument, the results may not reflect their true knowledge or attitudes; this may be due to familiarity with the content. Finally, some external validity is compromised, since the results are generalizable only to those who undergo such repeated testing.

Another set of issues arises from the use of longitudinal designs. In some cases, different results have been obtained when using a longitudinal design (which tests the same people over time) from those obtained when using a cross-sectional design (which tests different people at different stages at the same time). When considering such designs, you may want to consult literature on this topic (Russ-Eft 1999; Russ-Eft and Hoover 2005; Schaie, Campbell, Meredith, and Rawlings 1988).

Advantages of the time-series design:

- Controls for the effects of history
- Provides evidence over time
- Establishes baseline of performance with which to compare postintervention performance

Disadvantages of the time-series design:

- May require more resources because of repeated data collection efforts
- May be expensive because of repeated data collection efforts
- Changes may be the results of the repeated data collection (because of testing or instrumentation effects)
- May experience attrition or loss of people from the sample

Pretest-Posttest Control-Group Design

This design requires two groups, one that participates in the learning, performance, or change intervention and one that does not. Again, random assignment to the two groups provides greater internal validity to the study but may not be feasible. This design is depicted in Illustration 6.7.

Let us return to Tim's evaluation. To use this design he would first randomly assign salespeople to two groups, one of which would participate

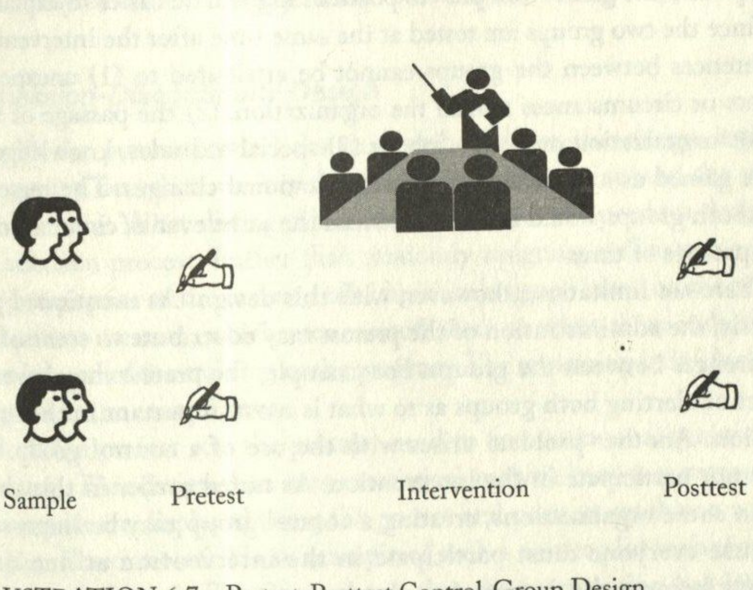


ILLUSTRATION 6.7 Pretest-Posttest Control-Group Design

in the training first. Each group would complete a pretest at the same time, such as a test on sales procedures. After the first group completed the training, the two groups would receive a posttest at the same time, again on sales procedures.

This design is one of the most rigorous described so far. The random assignment to the two groups eliminates the possibility of any selection bias. The pretest allows the evaluator to determine empirically whether the two groups are similar before training. If random selection is used, then the groups will likely appear similar before training. In the case described in the paragraph above showing some selection bias, differences will appear in the pretest. For example, if new salespeople were chosen for the group receiving the intervention and the experienced salespeople were chosen for the control group, a knowledge pretest would show that the new salespeople were less knowledgeable than the experienced salespeople. After training, the new salespeople may have gained some knowledge, but they may still receive lower scores than the experienced salespeople. In this case, the decision makers may view the training as unsuccessful, when in fact it may have been highly successful. By using comparable groups before training, the posttraining differences between the groups and the gains from pre- to posttraining will be easier to explain.

Since the two groups are tested at the same time after the intervention, differences between the groups cannot be attributed to (1) unexpected events or circumstances within the organization, (2) the passage of time in the organization or on the job, or (3) special attitudes, knowledge, or skills gained as a result of certain organizational changes. The reason is that both groups would have experienced the same events, circumstances, and passage of time.

There are limitations, however, with this design. As mentioned previously, the administration of the pretest may contribute to some of the differences between the groups. For example, the pretest may have the effect of alerting both groups as to what is most important in the intervention. Another problem arises with the use of a control group that does not participate in the intervention. As noted earlier in this chapter, in some organizations, creating a control group may be impossible because everyone must participate in the intervention at one time. Finally, because the design takes place over time, one must expect some attrition from the two groups, given the mobility of today's workforce.

If attrition from the two groups were unequal, then this would be a cause for concern.

Advantages of the pretest-posttest control group design:

- Provides some control for the effects of history and maturation because of the use of a control group
- Provides evidence over time
- Measures actual attitudes, knowledge, and skills prior to the intervention
- Allows for comparison of actual pretest and posttest data

Disadvantages of the pretest-posttest control group design:

- May require more resources because of repeated data collection efforts
- Changes may be the result of the repeated data collection (because of testing or instrumentation effects)
- May experience attrition or loss of people from the sample
- May have groups that are not similar because of unequal attrition
- May be difficult or impossible to obtain a control group

Regression-Discontinuity Design

As with the pretest-posttest design described above, this design requires two groups, one that receives the training or intervention and the other that does not. What distinguishes this design from the previous design is the selection process. Rather than randomly assigning individuals to the two groups prior to the pretest, individuals are assigned based on the results of a pretest measure, which uses a cutoff score to select those individuals most in need of the training or the program. The remaining individuals are assigned to the control group.

Both the pretest and the posttest need to be continuous, quantitative measures. This design then uses a statistical analysis known as regression. The results are displayed using the pretest and posttest results on a scatter plot. Illustration 6.9 is an example of such a scatter plot with the cutoff score indicated. The figure shows that the intervention has had no effect, since there is a straight line:

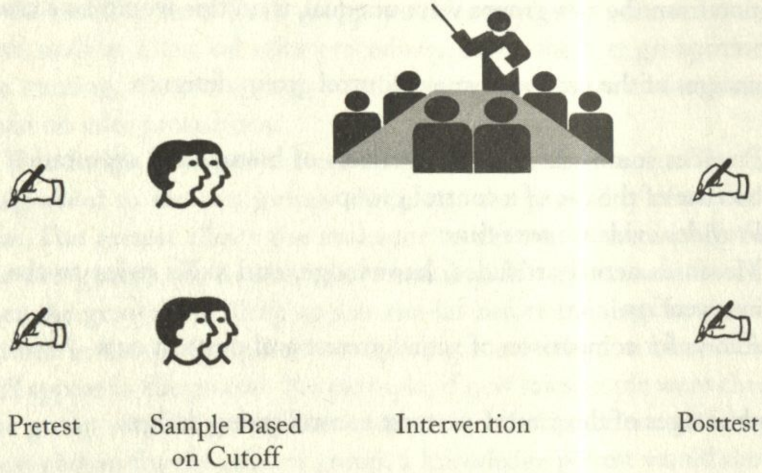


ILLUSTRATION 6.8 Regression-Discontinuity Design

Next, let us assume that training has taken place for the group most in need and not for the control group. Let us assume that the intervention was effective and it raised the posttest scores of the treatment group. Following the posttest, the regression lines are plotted separately for the treatment and the control group. The scatter plot in Illustration 6.10 shows a discontinuity in the regression lines. This discontinuity indicates that the treatment had a positive effect, since the treatment group scores have risen and are offset from those of the control group. In fact, the figure shows that the treatment group scores are ten points higher, on average, than those of the control group.

Evaluators have begun to examine the efficacy of the regression-discontinuity design and have found that it demonstrates high internal validity. At the same time, it enables program administrators and evaluators to assign people according to their need for the intervention rather than randomly. More about this design and its use can be found at the Research Methods Knowledge Base (www.socialresearchmethods.net/kb/index.php), with specifics about the design (www.socialresearchmethods.net/kb/quasird.php), as well as at the Campbell Collaboration website (www.campbellcollaboration.org).

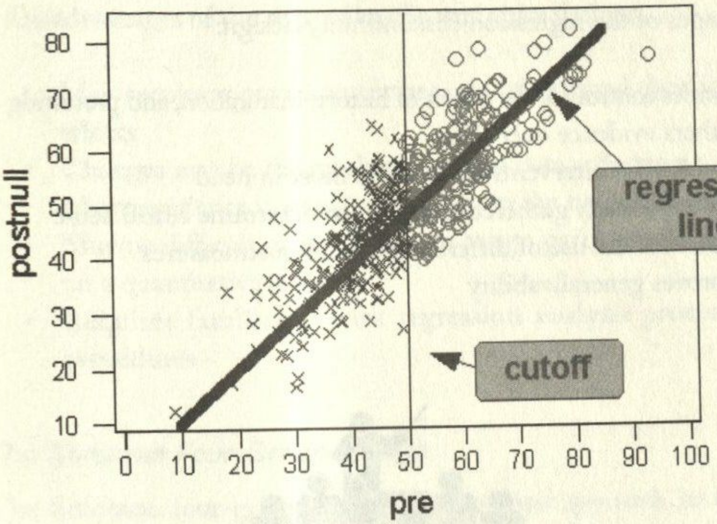


ILLUSTRATION 6.9 Pre-Post Distribution with No Treatment Effect

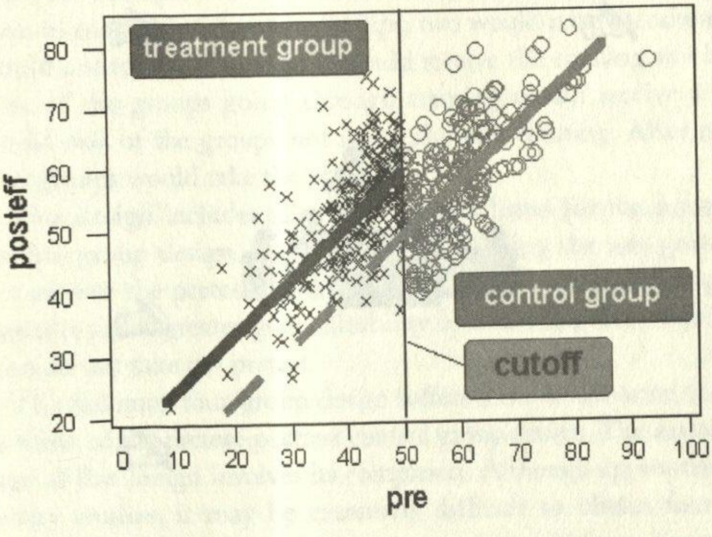


ILLUSTRATION 6.10 Regression-Discontinuity Design with Ten-point Treatment Effect

Advantages of the regression-discontinuity design:

- Provides control for the effects of history, maturation, and pretesting
- Gathers evidence over time
- Provides the intervention for those most in need
- Can use already gathered measures to determine cutoff score
- Allows for the use of different pre- and postmeasures
- Improves generalizability

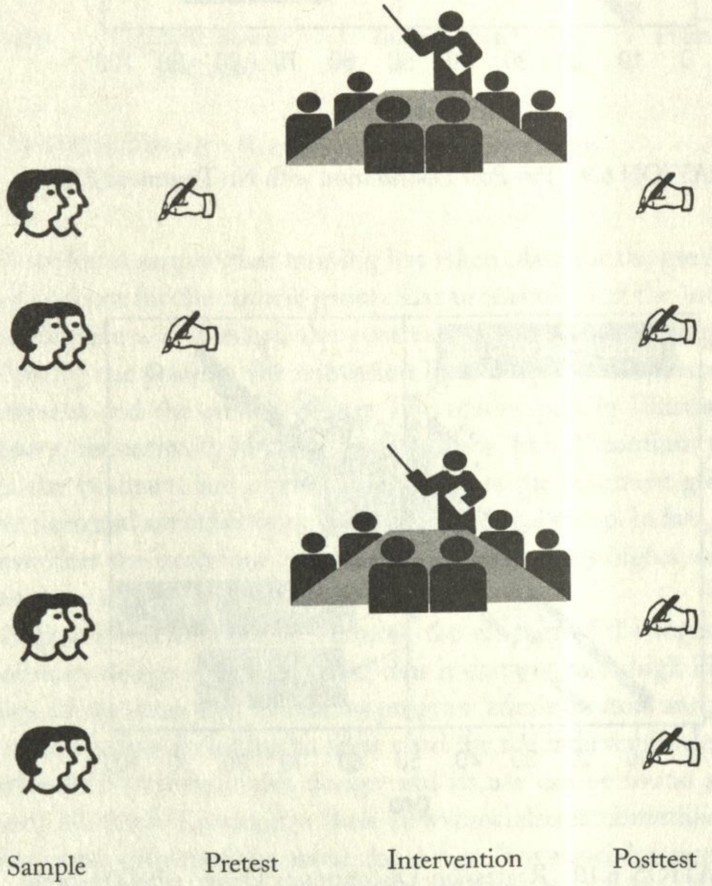


ILLUSTRATION 6.11 The Solomon Four-Group Design

Disadvantages of the regression-discontinuity design:

- May require more resources because of repeated data collection efforts
- Changes may be the result of repeated data collection
- May experience unequal attrition from the two samples
- May be difficult or impossible to assign participants only based on a quantitative measure
- Requires familiarity with regression analysis processes and procedures

The Solomon Four-Group Design

The Solomon four-group design provides one approach to controlling the effects of pretesting. It is probably the most rigorous design in terms of its ability to control for internal and external threats to validity. In this design, four groups are randomly selected and measured as shown in Illustration 6.11.

In our example with Tim and the sales training program, Tim would have to randomly select four groups; two would receive training and two would not receive training (or would receive the training at a later time). One of the groups going through training would receive a pretest, as would one of the groups not going through training. After training, all four groups would take the posttest.

This design includes all of the benefits listed for the pretest-posttest control group design. In addition, by including the two groups that do not receive the pretest, it controls for any effects of pretesting. Furthermore, it yields greater generalizability by extending the results to groups who do not take the pretest.

The Solomon four-group design suffers some of the same disadvantages as those of the pretest-posttest control group design. The major disadvantage of this design involves its complexity. Although appropriate for laboratory studies, it may be extremely difficult to obtain four randomly assigned groups and measure them over time within an organization. In particular, it will be difficult to have two groups that are randomly selected not receive the intervention. Also, because the design takes place over time, some unequal attrition from the groups will likely occur and may destroy the assumptions of random selection. Finally, with four groups, it requires a lot of people, a lot of administration, and a lot of data collection.

Advantages of the Solomon four-group design:

- Provides control for the effects of history, maturation, and pretesting
- Provides evidence over time
- Improves generalizability

Disadvantages of the Solomon four-group design:

- May require more resources because of repeated data collection efforts
- May be expensive because of repeated data collection efforts
- Changes may be the result of repeated data collection
- May experience unequal attrition from the sample, resulting in four groups that are not comparable
- May be difficult or impossible to obtain two control groups that do not receive the intervention
- Requires a large sample

The Case Study Design

Case studies involve in-depth descriptive data collection and analysis of individuals, groups, systems, processes, or organizations. The case study design is most useful when you want to answer how and why questions, and when there is a need to understand the particulars, uniqueness, and diversity of the case (Stake 1995; Yin 2003a,b). Case studies typically employ qualitative methods such as individual and focus group interviews, observation, and archival records, though they frequently include quantitative data from surveys or tests. Qualitative methods are emphasized since they are more effective in uncovering individuals' attitudes, beliefs, assumptions, values, practices, and experiences. For example, within the context of learning, performance, and change, we could better understand (1) how employees are using what they learned from a training event, (2) why employee performance is at a certain level, (3) the ways in which individuals learn from technology-based instruction, (4) how the organization supports or inhibits individual, team, and organizational learning and change, or (5) what effect reorganization had on a certain group of employees.

A case study design is particularly useful when the evaluator has little or no control over events, and when it is important to study organization members within their natural setting (the organization). The evaluator using a case study design does not seek to control or manipulate the environment or any of its variables. Instead, the evaluation focuses on "what is" and tries to explain or make sense of the phenomenon being evaluated according to those who experience it. In evaluating learning, performance, and change initiatives, the goal would be to construct a holistic understanding or gestalt of the organization members' context. To understand the various meanings associated with learning, performance, or change as a result of some intervention or initiative, case studies might use multiple methods and sources of data that are collected over time (this might be one day or several months). Case studies are most effective when the reader of the case vicariously experiences what it might have been like to be there. Thus case study designs are particularly appropriate and useful when:

- Context is of critical importance
- Understanding is the primary goal
- Multiple sources of evidence are sought
- The evaluation questions focus on the "how" and "why" of something
- Generalization of findings is not the primary goal

Tim's evaluation of the sales training program could have used the case study method instead of the one-time test. Such a study could have gathered in-depth background information on the trainees, followed by observations and interviews of trainees before, during, and after the training.

Advantages of the case study design:

- Provides descriptive data
- Does not require manipulation or control of individuals or the setting
- Reports include verbatim quotes of those interviewed

- Leads to greater understanding about the context of the evaluand
- May lead to greater understanding about practice
- Tends to gather data using multiple methods (triangulation)
- Provides data that are rich with examples and stories
- Captures what is important to participants
- Portrays the multiplicity of causes associated with various outcomes
- Embraces the diversity of perspectives and experiences of participants
- Allows the evaluator to collect information on outcomes not known or anticipated prior to the learning and performance initiative

Disadvantages of the case study design:

- Results do not lead to scientific generalizability
- Evaluator bias may interfere with validity of findings
- May take too long to conduct
- May produce more data than can be analyzed in a timely or cost-effective way

Once you have decided on the most appropriate evaluation design or designs, you need to describe them in your evaluation plan. Now you can focus on which methods will collect the necessary data.

The Mixed Methods Design

Rather than relying solely on a qualitative case study or solely on an experimental design, evaluators are increasingly using mixed methods. Such designs gather both qualitative and quantitative data in order to answer the evaluation key questions. Tashakkori and Creswell (2007) define mixed methods as "research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry" (p. 4). A mixed method design allows the evaluator to examine questions such as What? How many? and To what extent? coming from the quantitative approach, as well as Why? questions coming from the qualitative approach. In addition, employing both approaches can help decision makers get the hard numbers they need, while enabling the evaluation to interact with the participants and to experience the culture. As stated by Willging, Helitzer, and Thompson (2006), "rather than focus on culture as a static, homogenized entity, it is more fruitful . . . to focus on the intended audience's interactions within broader social and physical environments" (p. 138).

Keep in Mind . . .

- If possible, evaluations should include designs that collect both quantitative and qualitative data.
- Evaluation team members should address issues related to both internal and external validity.
- The choice of one or more evaluation designs should reflect the evaluation's key questions.