CHAPTER ONE

EXPERIMENTAL AND QUASI-EXPERIMENTAL EVALUATIONS

S. Bartholomew Craig and Kelly M. Hannum

Experimental and quasi-experimental approaches to evaluation are the focus of this chapter, and they provide a structured means to think about designing evaluations. Though leadership development initiatives and quasi-experimental designs have both been around for decades, few published resources address the challenges of applying experimental or quasi-experimental designs to leadership development.

Two challenges faced by many, if not all, evaluators of leadership development initiatives are (1) the need to measure changes in leadership or leadership outcomes—two complex and sometimes nebulous areas; and (2) determining the relationship between the leadership development initiative in question and the changes measured. Experimental and quasi-experimental approaches provide a means to address both challenges.

Research and evaluation can be thought of as distinct but related activities. This chapter represents the overlap between research and evaluation and uses a research framework for thinking about evaluation. Research designs are typically categorized in one of three ways: nonexperimental, experimental, or quasi-experimental. Exhibit 1.1 illustrates the key elements of each design. Nonexperimental designs are observations about something without any intervention in what is being studied. Because there is no intervention in

EXHIBIT 1.1. USEFUL TERMS AND DEFINITIONS.

Control Groups. A control group contains people who did not participate in the initiative being studied. This is the group against which data from those who did participate in the initiative are compared.

Random Placement. Individuals are assigned to participate in a program or not through a random method and not on the basis of any characteristic they possess or any other non-random process.

Nonexperimental Design. Observations are made in the absence of any intervention in the phenomena being studied. Relative to the other designs, nonexperimental designs are comparatively inexpensive and simple to execute, but provide only hints about possible cause-and-effect relationships. *Example:* Examining the relationship or correlation between leaders' uses of rewards and team performance, without making any attempt to influence either, would be an example of a nonexperimental design.

Quasi-Experimental Design. Observations are made about an intervention. Control groups are typically used, but groups are not created using random assignment. These designs are more complicated to implement than nonexperimental designs, but provide more information about possible cause-and-effect relationships. *Example:* Leaders choose whether to participate in a leadership development program or not. Those participating in the program are compared to themselves before the program or to other groups who did not participate in the program.

Experimental Design. Observations are made about an intervention. One or more control groups are used, but groups are created using random assignment. Because random placement reduces the need to prove that groups are roughly equivalent, results from these designs can be less complicated to interpret than those from quasi-experimental designs and can provide the most information about possible cause-and-effect relationships. *Example:* A group of leaders is identified as appropriate for a leadership development program. The group is randomly divided into two cohorts with one group participating in the leadership development program first. Those participating in the program are compared to the group that has not yet participated in the program.

The table following provides an overview of the key distinctions among the three approaches to research addressed in this chapter.

	Nonexperimental	Quasi-Experimental	Experimental
Control Groups	No	Usually	Yes
Random Placement	No	No	Yes

nonexperimental studies, no evaluation would be considered nonexperimental. Experimental and quasi-experimental designs involve interventions of some kind. Almost all program evaluations could be considered to be at least quasi-experimental in nature, because the program being evaluated represents an intervention that would not have occurred otherwise. In the context of evaluation, however, the terms *experimental* and *quasi-experimental* usually imply that data from different groups are to be compared in some way. This comparison may be made across time, as when the same participants are assessed before a leadership development program and then again afterward; or, the comparison may be made across people, such as when managers who participated in a development program are compared to managers who did not.

When comparisons are made among groups of people, the distinction between experimental designs and quasi-experimental designs comes into play. In experimental designs, individuals are randomly assigned to participate in programs. The group not participating in the program is usually called a *control group.* The control group is the group against which those participating in the program are compared. In quasi-experimental designs, individuals are put into groups on the basis of some nonrandom factor. For example, if leaders are allowed to choose whether or not to participate in a given program, then any evaluation of that program would be, at best, quasi-experimental, because participants were not randomly assigned to participate.

Random assignment is the reason why experimental designs can be more effective at establishing whether the program actually caused the changes that were found. Randomly assigning participants into a program (or not) allows evaluators to assume that any preexisting differences among individuals are evenly distributed between the group participating in the program and those in the control group. For example, some individuals are more ambitious than others. If individuals are allowed to decide for themselves whether to participate in a leadership development program, a greater number of ambitious individuals may become participants versus individuals who are not ambitious. If the evaluation later finds that program participants tended to rise to higher levels in the organization than did nonparticipants, there would be no way to know whether that difference existed because of the program or because the participant group was more ambitious and therefore engaged in other processes that furthered their careers. By randomly assigning people to participate in leadership development, the evaluator can assume that program

participants are no more ambitious, on average, than control group members. A similar issue arises when supervisors are asked to recommend individuals for program participation; they may be more likely to recommend high performers in order to maximize return on the organization's investment in the program. The list of factors that can influence group membership in quasi-experimental designs is nearly endless.

If the evaluator could anticipate all the factors that might influence whether individuals end up participating in the program, it would be a simple matter to compare the group participating in the program to those who did not participate in the program, because the evaluator could statistically account or control for any differences that existed prior to the program. Unfortunately, there is no way for an evaluator to be certain that all possible variables have been considered. The main strength of random assignment is that the evaluator is relieved of that burden. When people are assigned to groups at random, individual differences can be assumed to be evenly distributed across the groups (for example, intelligence, personality, work experience, sex, and race). The goal of an experimental design is to engineer a situation in which the only difference between the groups being compared is participation or nonparticipation in the program. Even with randomization it is possible, though unlikely, to get nonequivalent groups; you have experienced this if you have ever tossed a coin and had it come up heads far more often than tails. Therefore, it is recommended that randomly assigned groups be spot-checked on a few key variables of interest in order to confirm that randomization succeeded. If the groups are found to be nonequivalent, randomization can be repeated or the variables on which they differ can be controlled for with statistical techniques. When experimental designs are impractical, as is often the case in the leadership development context, quasi-experimental designs may be used to increase confidence regarding causality.

We began this chapter with two challenges faced by evaluators of leadership development initiatives. The two challenges are related to two broad questions, the answers to which are often sought as part of evaluations. The first question is, What changes have occurred? Answers to this question might include the specific domains where change was found (for example, self-awareness), the direction of the change (for example, increase or decrease), the magnitude of the change (for example, by 21 percent), and the level at which the change occurred (for example, individual, team, or organizational). In most cases quasi-experimental and

experimental evaluations of leadership development programs are concerned with change reported at the individual or team level. (It would be a difficult but illuminating study to look at organizations conducting leadership development and compare them to organizations not conducting leadership development.)

The second broad question is, Were the changes caused by the program being evaluated? Change can occur for a variety of reasons. It is usually not sufficient for an evaluation to describe how people changed around the time of a leadership development program; we want to know whether that change occurred *because* of the program. Summative evaluations focus primarily on those two main questions. Formative evaluations also consider questions specifically related to the process or functioning of an initiative with an eye toward how it might be improved.

This chapter is organized around the two questions just mentioned. First, we address some issues associated with measuring change. We then proceed to the problem of linking the changes found to the leadership development initiative. Evaluations are often conducted to help make decisions about resources. If the desired changes are not happening or are not the result of a program, then changes to the program or an entirely different program might be needed. Those using information from evaluations must balance the risk of altering or dropping a program that is performing well, even though the evaluation is not able to pick up on the benefits of the program, with the risk of continuing to put resources into a program that may not be delivering results despite the positive spin around the program. Before delving into these two evaluation questions, we first consider the context in which the evaluation is to occur. The context is critical because it offers clues into what type of evaluation is most appropriate.

Evaluation Context

Leadership development programs, and evaluations of them, are conducted in a wide variety of settings for a wide variety of purposes. The specific context in which the evaluation is to take place has implications for whether or not an experimental or quasi-experimental design is appropriate and, if so, what specific design may be most suitable. Several aspects of the context that should be considered are discussed next.

Clarity of Objectives and Outcomes

All too often, leadership development initiatives are implemented without a clearly stated set of objectives or outcomes. Goals for the program might be stated in vague terms, such as "improve leadership capacity" or "develop our leadership pipeline" with no specific objectives or outcomes associated with the goals. In such cases, different stakeholders may have different ideas about what the program is supposed to accomplish because the goals are open to interpretation. Part of an evaluator's role is to help ensure that stakeholders have a shared understanding of the program's objectives and outcomes. Chapter Two provides helpful advice about how to facilitate and document conversations about stakeholders' expectations for an initiative. This clarification provides necessary information about what changes are expected to occur. Most leadership development initiatives are expected to cause change in several areas, such as participant self-awareness, interpersonal skills, or approaches to problem solving. The direction in which the change should occur should also be clarified with stakeholders. For example, self-awareness might increase, decrease, or stay the same. In almost all cases, measurement strategies should be selected that are capable of detecting change in any direction.

In addition to creating confusion among stakeholders, vaguely stated goals are difficult to measure. Before specific measures can be selected or developed, desired objectives and outcomes must be stated in unambiguous language. Ideally, this clarity should be attained early in the design of the leadership development initiative. Several experimental and quasi-experimental evaluation designs involve collecting data before the initiative begins (pretests). If the desired objectives and their outcomes are not articulated until after the initiative has begun, then pretests are less likely to be useful or may not be possible at all. An initial evaluation may be needed to gather qualitative evidence about what changes occur that can later be used to develop or select more targeted measures.

In situations where stakeholders are not clear about the changes they expect to see after a leadership development initiative, an experimental or quasi-experimental design may not be the best approach. It may be wise to begin first with a more flexible approach that can help deepen understanding about what changes are occurring after an initiative, the results of which could be used to help design an experimental or quasi-experimental evaluation. Chapter Three in this book describes an open-systems approach that may be helpful in contexts that may not be appropriate for experimental or quasi-experimental designs.

Availability of Sound Measures

Even when stakeholders have clearly stated their objectives and outcomes for a leadership development initiative, some of the desired objectives and outcomes may not be easily measurable. For instance, a program goal may be to improve participants' ability to adapt to a changing competitive landscape. But exactly how you measure "ability to adapt to a changing competitive landscape" may be far less clear. How would we quantify this dimension so that participants could be compared, either to themselves before the program or to a control group?

In addition to measuring specific changes, it is important to consider the amount of detail needed about "how much" change has occurred. For instance, you may want to administer a measure of self-awareness that could detect improvement by 5 percent or by 75 percent, or it may be enough to know that, in general, there were signs of improvement. There is little point in using an 18-point response scale when all that is needed to address stakeholder questions is a simple Yes or No.

In some cases, an objective or outcome may be so specific to a particular organizational context that established measures of it do not exist. When outcome criteria are difficult to quantify, evaluation designs that require comparisons among groups will be difficult to implement. Certainly, objective performance data such as revenue or employee turnover rate lend themselves to quantitative comparisons, but there may not be a strong enough logical link between changes in these measures and the leadership development initiative being evaluated. Sometimes individuals will want to make comparisons among data that are readily available without using sound logic to link the measure and the initiative. When thinking about what kind of measure to use, it is important to be sure that what you want to know about can in fact be measured relatively accurately and that there is a reason to think that the initiative will have a fairly direct impact on what is being measured.

Availability of Adequate Sample Size

Experimental and quasi-experimental designs involve comparisons, typically conducted using statistics that explain whether the differences between the groups are likely to be chance fluctuations or real impact. For comparisons to be defensible, fairly large sample sizes may be needed. If a leadership development

initiative is only implemented with a small number of individuals (twenty, for example), such comparisons may not be statistically viable (Tourangeau, 2004). If there is not an adequate number of people participating in the program to conduct statistically meaningful comparisons, it does not make sense to invest the time and money into an experimental or a quasi-experimental design. Some resources for determining sample size requirements are listed at the end of this chapter.

Initiative Time Span

One of the key ways in which evaluation designs differ from each other is in terms of the timing of data collection. For example, pretests typically are thought of as occurring before an initiative starts and posttests as occurring after the initiative ends. But some leadership development initiatives may not have definite beginning or end dates. This is often true in the case of systemic leadership development initiatives. Systemic approaches to leadership development may involve a sequence of developmental job assignments or mentoring relationships that are ongoing, with no specific end date. Leaders participating in this kind of development usually do not move through the system as an intact group or cohort; different individuals are at different stages at any given point in time. Evaluations of such initiatives cannot wait for the program to be completed; evaluators must employ designs that collect data at meaningful time points that may be different for different participants, which increases the complexity and complicates the interpretability of experimental and quasi-experimental designs.

Environmental Stability

One of the most important and challenging aspects of leadership development evaluation is establishing the program as the cause of the observed changes. People may change for a variety of reasons that have nothing to do with participation in the program being evaluated. Changes in the organizational context can lead to changes in individuals. For example, if the goal of an initiative were to increase participants' willingness to take risks, and the organization underwent a merger during the course of the program that caused some managers to fear losing their jobs, measures of "risk taking" taken after the program might not accurately reflect the program's efficacy in that do-

main. In fact, the environmental event (the merger) might decrease the apparent effectiveness of the program by causing it to appear that participants were actually more risk averse after attending the program. Other environmental events that could produce similar results include organizational restructuring, changes in organizational leadership, changes in funding or budget allocations, the entry of new competitors into a market, the introduction of new policies and procedures, new rewards and recognition systems, changes in the regulatory or legal landscape, and changes in the political regime of the country in which the organization operates. The list of possibilities is extremely long and highly dependent on the context in which the evaluation is being conducted. Ideally, evaluations should be timed so as to be as insulated as possible from potentially disruptive environmental events. When evaluations must take place in unstable environments, evaluators should make careful note of the relative timing of the events. When possible, evaluators should also take separate measurements of the events' effects to have the best possible chance of being able to separate their effects from those of the program. In unstable environments, using control groups who experienced the same environment as participants but who did not participate in the program is especially useful.

Measuring Change

A critical part of the design process is deciding what kinds of change will be measured and how. Measuring leadership outcomes and linking them to a specific initiative in dynamic and fluid contexts is by no means simple. Ideally an evaluator would work with stakeholders to determine the areas in which change can be expected and linked to the leadership development initiative and to determine how the change can best be measured. Once the domains are identified, appropriate and accurate measures for assessing that domain can be identified or developed. For example, an evaluator may decide to measure participants' self-awareness by comparing self and others' ratings on a 360-degree assessment instrument, or the evaluator might interview participants' colleagues to ask how effectively participants communicate their visions for the future to others. As part of this process it is also important to identify the level(s) at which change is expected (for example, individual, group, organizational, community), when the change is expected, and from whose perspectives the change can be seen and measured.

It may seem obvious, but it is critical to be certain that the measures you are using are as accurate and as appropriate as possible. In many cases, positive behavioral change is an expected outcome of leadership development. Accurately measuring behavioral change is difficult and much has been written about this topic (for example, Collins and Sayer, 2001; Gottman, 1995; Harris, 1963). Relying on instruments with established, well-researched psychometric characteristics is one way to help ensure accurate and appropriate measures. The two indicators of most interest are reliability and validity. They are related concepts, but they have distinct meanings and are assessed differently.

The reliability of a measure can be thought of as the consistency of results (see Exhibit 1.2). For example, if a scale indicates you weigh 120 pounds on one day and then on the next day it indicates you weigh 220 pounds, those results are inconsistent, which means the scale is not a reliable measure of your weight. Reliability can be estimated a number of ways and is usually indicated on a scale from zero to one. Typically, reliability estimates above 0.80 are considered reasonably good (Nunnally, 1978). However, it is important to keep in mind that some things can be measured more objectively (for example, the frequency with which a manager provides feedback) while other areas can only be measured subjectively (for example, the quality of the feedback provided). Objective measures are more likely to have higher reliability estimates. Reliability is important because if a measure is providing inconsistent results, you may not want to put too much stock in the data you collect with it.

When using a preexisting measure, it is also important to make certain that the measure is a good fit for the situation, which leads us to the appropriateness of the measure or the measure's *validity*. A measure of coaching behaviors developed for use with sports team coaches is not likely to be a good measure for the coaching learned in a leadership development program for a manufacturing setting. Similarly, an assessment developed for use with university students in Sweden may not be appropriate for university students in Venezuela. There are various approaches to determining a measure's validity. Exhibit 1.2 provides an overview of some of the more common approaches to validity.

Cause and Effect

Whether or not, or how confidently, the second basic question of evaluation—Was the change caused by the program?—can be answered depends on the design of the evaluation. Typically an evaluation design provides a logical plan

EXHIBIT 1.2. RELIABILITY AND VALIDITY OF A MEASURE.

Reliability is the degree to which an assessment produces consistent results. If an assessment does not produce consistent scores, you may be getting more error than information. Reliability is never truly measured, but it can be estimated. The same test will likely have different reliability estimates depending on how reliability is calculated and on the sample used. The appropriate reliability level depends on the situation. Reliability is usually reported on a scale ranging from 0 to 1, with estimates closer to one being preferred. Three ways to assess reliability are

1. *Internal consistency,* which provides information about whether items on a scale are measuring the same or closely related concepts. Usually Cronbach's alpha is used to measure internal consistency. The Instrument Review Team at the Center for Creative Leadership, for example, recommends alphas of 0.70 or higher.
2. *Interrater agreement,* which provides information about the degree to which ratings agree. *Feedback to Managers* suggests interrater reliabilities should be between 0.40 and 0.70 for 360-degree assessments (Leslie and Fleenor, 1998).
3. *Test-retest,* which provides information about the stability of items and scales over time. In this case, the test is administered and then administered again after a short period of time. Reliabilities of 0.70 or higher are generally considered acceptable.

The validity of a test is a combination of two ideas: (1) the degree to which an assessment measures what it claims to measure; and (2) the usefulness of an assessment for a given purpose. Validity is a multifaceted concept and an extremely important consideration when developing or using assessments. Multiple types of evidence are needed to establish test validity. Validity evidence should be gathered in the varying situations and with the varying populations for which the assessment is intended. Validity has to do with the test, the people taking the test, the purpose of the test, and the consequences of the test. Types of validity evidence for assessments include:

- *Content validity.* The extent to which the assessment adequately and comprehensively measures what it claims to measure.
- *Construct validity.* The relationship between test content and the construct it is intended to measure. Typically, this type of evidence involves logical and/or empirical analysis including statistical comparisons to other assessments and expert judgments of the relationship between the assessment and the construct.
- *Criterion validity.* The relationship between the assessment and a criterion such as effective performance (for example, looking at the relationship between an assessment of job performance and job performance ratings). Concurrent evidence refers to criterion data collected at the same time the test is administered and predictive evidence involves criteria collected at a later point in time.

EXHIBIT 1.2. RELIABILITY AND VALIDITY OF A MEASURE, Cont'd.

- *Consequential validity.* Evidence supporting the benefits and consequences of testing are examined in this type of study. Consequences of tests are considered aspects of validity when they are related to construct underrepresentation or construct-irrelevant components of a test. This is particularly important in high-stakes testing. The consequences associated with test use are not universally accepted as an aspect of validity.

We sometimes take for granted that an assessment is providing accurate, useful, and appropriate information. Assessments do not always do that. Validity studies are one way that item or test bias or unfairness can be revealed. Bias is the presence of an item or test characteristic that results in differential performance for individuals of the same ability but from different groups (for example, ethnic, sex, cultural, social status, or religious groups). Bias often stems from limitations of our perspective and understanding. No test is free from bias, but item and test bias and unfairness can be detected and reduced. How might bias enter into an assessment? Items that use vocabulary, content, structure, or an administration mode that improve the performance of one group over another are potentially biased or unfair items. For example, a test written in English might be biased against individuals for whom English is a second language. Other potential sources for bias or unfairness include offensive, demeaning, or emotionally charged items. While there are strategies and tools for assessing bias in tests and items, it can be difficult to figure out how much difference is too much, the root of the difference, and the appropriate course of action. The process can get complicated and expensive. The consequences of not addressing bias in assessments can also be complicated and expensive. Assessments are only as good as we make them. How accurate an assessment needs to be depends on many things including the intended use and consequences of use associated with the assessment.

for what will be assessed, how it will be assessed, when it will be assessed, and from what sources data will be collected. Linking changes in leadership outcomes to a leadership development program cannot be accomplished without a good evaluation plan. Drawing conclusions about cause and effect is almost never straightforward. A general discussion about causal inferences is included in the overview for this part of the book. How confident we can be that the changes we measured can be attributed to the leadership development program in question depends heavily on how the evaluation was designed.

Factors that reduce our confidence about causality are called *threats to validity.* Threats to validity are possible alternative explanations, not related to the program, about why changes may have been observed. For instance, if par-

ticipants received a large monetary bonus during the time period when the leadership development program occurred, that—rather than the program—might be the reason for any increase in organizational commitment. As we discuss in detail later, different types of evaluation designs are vulnerable to different types of threats to validity. Understanding validity and threats to validity is essential for making methodological choices.

Validity

Validity is the truth of inferences based on the results of your evaluation. Validity, as we discuss it in this section, is about the evaluation rather than a specific measure (see Exhibit 1.2). Strong validity requires accurate, appropriate, and sufficient evidence. An evaluation is said to have adequate internal validity if we can be confident in its conclusions about cause-and-effect relationships. For example, if an evaluation provides compelling evidence that managers' participation in a leadership development program caused their sales teams to increase their orders in the month following the program, then the evaluation study would demonstrate strong internal validity. What is considered compelling evidence is a matter of judgment. In order for us to have confidence in such a cause-effect relationship, the study's design must enable us to rule out other plausible explanations for the increased sales, such as seasonal fluctuations or a broad change in market demand. In many ways the situation is similar to a legal argument; is there convincing evidence that the program caused or contributed to the changes indicated? It is important to consider that different stakeholder groups may have very different ideas about what they consider convincing evidence and may be able to offer differing perspectives on the logic of arguments.

External validity is the degree to which conclusions from an evaluation are true for people, places, or times other than the ones actually evaluated. For instance, if another evaluation conducted a year later on a different group of participants found the same effect on sales orders, we would say that the first study demonstrated external validity. An evaluation study must have internal validity in order to have external validity, but having internal validity does not guarantee external validity.

Threats to Internal Validity

Factors that weaken confidence in conclusions that changes were caused by the program being evaluated are called *threats to internal validity.* As mentioned

earlier, different evaluation designs are vulnerable to different threats, so an understanding of common threats to internal validity is important to anyone charged with designing or interpreting an evaluation study.

Systematic Differences Between Program Participants and Nonparticipants. As mentioned earlier, a key goal of an experimental or quasi-experimental design is to create a situation where the only difference between program participants and nonparticipants is their participation in the program. This is an almost impossible goal in today's dynamic environments, but the closer you can get to it, the more confidence you can have that the changes observed are because of the program evaluated. This is generally achieved by randomly assigning individuals to the program (experimental) or measuring and controlling for factors that may be different between participants and nonparticipants (quasi-experimental). Several potential threats to achieving such a state of affairs are explained below.

Selection. Any factor that causes people with certain characteristics to be more likely to participate in the program than people without those characteristics is a threat to internal validity. For leadership development initiatives, two common practices are self-selection and boss-selection. If participants are permitted to choose whether to participate (self-selection) or their superiors select or nominate them for participation (boss-selection), then people with certain characteristics may be more likely to be excluded from the program group (for example, managers with more hectic schedules, lower ambition, lower or higher job performance). In cases where self-selection or boss-selection is used to identify who will participate in leadership development, randomly creating two cohorts allows for a control group that is likely to have similar characteristics. For example, if fifty individuals were selected by themselves or their bosses to participate in leadership development then twenty-five could be randomly selected to participate in the first cohort, while the remaining twenty-five could be used as a control group until their participation in the program.

Other potential problems include conducting the program at a time when certain types of people are not available or in locations that are more accessible to some types of leaders than others (for example, conducting a program during a time when individuals from a specific region are involved in opening a new office). Such factors can result in *preprogram* differences between groups that may later be confused with program effects. Selection can also be an issue

when not all program participants are included in the evaluation and the individuals participating in the evaluation are different from those who do not participate.

Attrition. This problem is similar to the selection issue except that attrition occurs when participants with certain characteristics are more likely to drop out of the program (or the evaluation) before its completion. The end result is the same as with selection: the groups being compared are different for reasons other than the program. For example, leaders who have difficulty delegating or who work in small organizations may be more likely to have to drop out of a program in order to deal with a crisis within their organizations. Attrition can occur in the context of the evaluation as well, with certain types of individuals dropping out of the evaluation (that is, failing to complete evaluation measures). For instance, individuals who did not experience benefits from the leadership development initiative may not want to spend more of their time by participating in the evaluation. If those who feel similarly also drop out, then the results of the evaluation are compromised. At a minimum, it is a good idea to follow up with individuals who drop out of the initiative or the evaluation to find out more about why they dropped out. In cases where a large number of people have dropped out, you may want to randomly follow up with a smaller subset of the individuals who dropped out.

Regression to the Mean. This issue concerns the tendency for those scoring extremely high or low on a measure to be less extreme during the next test. For example, if only those who scored poorly on a leadership capacities test are included in the program, they might do better on the next test regardless of the program just because the odds of doing as poorly the next time are low. Similarly, if only those scoring high on the leadership capacities test are selected, they may not do as well on the next test simply because achieving a higher score when one is already near the top of a range is difficult (there is less room for improvement). Selecting or developing a measure that more fully represents the knowledge, skills, abilities, or behaviors of those in the group targeted for development is one way to help guard against this threat.

Changes Not Caused by the Program. Whereas the threats discussed previously concern differences between groups that might mask change or be mistaken for change, the next two threats to validity exist when changes do occur

in program participants, but those changes are caused by some factor other than the program being evaluated.

Maturation. People are always changing and developing even when they are not explicitly involved in development programs. We tend to become better at our jobs with experience and our personalities can change with age. Distinguishing between changes resulting from natural biological or psychological development and changes caused by the program can sometimes be difficult, especially in evaluation designs that do not use control groups and continue over an extended period of time. Although maturation can be a threat in the evaluation of long-term systemic leadership development initiatives, this type of threat is less likely to occur in traditional leadership development situations of shorter duration (unless the program is specifically for new hires or new placements who might develop a variety of skills in a short time based on learning from their new job experiences). However, it can be a problem when the program is brief, but the time intervals between evaluation measurements are long, such as when pretests are conducted long before the program starts or posttests are administered long after it ends.

History. As discussed earlier, evaluations can be compromised when changes in participants' environment occur around the same time as the program and cause participants to change their behavior in ways that might be confused with effects of the program. Such events can occur in the internal environment of the organization or in its external environment. For example, a corporate merger could cause employees to fear for their jobs, or an economic recession might cause a downturn in organizational performance that could mask the otherwise positive effects of a leadership development program. Changes due to such historical events are more likely to be temporary than changes from maturation and can often be detected by collecting data at multiple points in time. Using a control group allows an evaluator to better tease out programmatic effects, since both those participating in the program and those in the control group should experience the same events and shifts.

Problems with Measures. Sometimes the evaluation process creates threats to its own validity. Although it can be useful to think of evaluation as a kind of intervention with its own set of outcomes, those outcomes can be problematic when they inhibit the evaluator's ability to accurately assess the effects

of the program. Two evaluator-generated threats to internal validity are discussed here.

Testing. You have probably heard the adage that "practice makes perfect." If identical measures are administered before and after a program, the preprogram test can act like a practice session that serves to improve performance on the postprogram test. This effect can occur for at least two reasons. One is familiarity; when participants take the test following the program, they have already seen it at least once before and that familiarity may serve to increase their scores the second time around. The other reason is that the pretest may provide participants with clues as to which parts of the program's content they should pay close attention to. By sensitizing participants to the specific areas on which they will be tested again after the program, the pretest serves as a study guide. In some ways this increased focus can be seen as a benefit, if knowing which areas are the focus of the program helps participants meet a learning objective. However, the result can be that the posttest scores become biased indicators of program effectiveness, usually overestimating how much participants learned in the program. If, however, the pretest is part of the initiative and there is no identical posttest that is part of the evaluation, testing is not likely to pose a threat to validity.

The testing threat is an issue primarily with knowledge or skill assessments that are completed by participants and contain items scored as right or wrong. This type of assessment is not very common in leadership development contexts, but it is used often enough that evaluators should be wary of the testing threat.

Instrumentation. Instrumentation bias occurs when measurements taken at different times are not meaningfully comparable. This can occur even when the measures have been taken with identical instruments. When instrumentation bias is operating, calculating the difference between scores from any two occasions can produce a misleading estimate of change. This can occur when instruments are completed by participants' coworkers, such as in the case of 360-degree leadership assessments. If the exact same group of coworkers is not surveyed at each time point, then differences in ratings may be due to the changes in the composition of the rater group rather than actual changes in participants.

Instrumentation bias can also occur in self-report instruments completed by participants. One reason for this phenomenon is often referred to as *response*

shift bias. Response shift can occur because the leadership development program changes the way participants think about the behaviors assessed by the instrument (Howard and Dailey, 1979; Rohs, 1999, 2002). For example, prior to the program a manager might rate herself as a "3" on a 5-point scale measuring her empowerment of her subordinates. During the program she learns about many new ways to empower subordinates that she had not considered before. When she rates herself again after the program, she realizes that she really should only have rated herself a "2" earlier, but because she has improved following the program she now rates herself a "3." Because both the pretest and posttest were 3s, a simple comparison of the two scores would suggest that no improvement occurred. But the change in the participant's frame of reference regarding what constituted a "2" versus a "3" prevents us from meaningfully comparing the scores. This change in frame of reference can be a positive outcome of the program, indicating that participants' knowledge in a particular domain has increased. But whether or not such a shift is considered to be a desirable outcome, it creates measurement problems and it is important to keep in mind that pre- and postprogram scores cannot be meaningfully compared when such an effect is present. Unfortunately, methods for detecting response shift bias involve sophisticated statistical procedures that are not accessible to all evaluators (for details, see Craig, Palus, and Rogolsky, 2000; Millsap and Hartog, 1988; Schmitt, 1982). As an alternative, some researchers have recommended the use of retrospective pretests that are administered at the same time as posttests in order to ensure that both measures are completed with the same frame of reference (Howard, 1980; Howard and Dailey, 1979). Retrospective measures require that individuals accurately remember behaviors exhibited in the past—frequently, a few months in the past. Thus retrospective measures depend on potentially faulty human memories, so there is at present no easy solution to the problem of response shift bias. In cases where different measures are used to assess the same construct or domain, then the two measures should be comparable in terms of content and response scale. This comparability is referred to as measurement equivalence and also requires sophisticated statistical techniques (Cronbach and Furby, 1970; Facteau and Craig, 2001).

Threats to External Validity

While an evaluation study's internal validity is a necessary requirement for external validity, it is not sufficient alone. *External validity* is the extent to which the evaluation's findings apply to people, places, or times other than the ones

actually studied. In some cases, evidence about the external validity of an evaluation study may not be critical. This is especially true when an organization wants the results only for use regarding a program in a specific context and for a specific group of individuals. However, in situations when program evaluation findings are intended to be reflective of results one might expect of a program across different types of individuals in different contexts, then guarding against the threats to external validity becomes important. In the paragraphs that follow we discuss some common threats to external validity that can occur even in a study with high internal validity.

Selection-Treatment Interaction. The pool from which participants are selected, even when they are randomly assigned to a program, can limit the generalizability of evaluation findings. For instance, assume a health care organization decided to evaluate its leadership development efforts intended for the high-potential leaders in the organization. A successful leadership development initiative for those participants, in that context, may not be effective for at-risk participants coming from the financial sector. Therefore it is important to consider the program within the context where it was implemented; for instance, in terms of the individuals participating, and the sector and region in which they are working.

Multiple Treatment Interference. In the context of leadership development, this type of threat can occur when participants have attended a previous development program (for example, a program on personal responsibility) and the effect of the prior program affects or interacts with the leadership development program. Future participants in the leadership development program who have not also experienced the personal responsibility program might exhibit different types of change than would have been expected based on the earlier evaluation. This situation limits the generalizability of evaluation findings because it is difficult to determine the effects due exclusively to the leadership development program. In cases where other initiatives are known to have been offered, the evaluator could list different development opportunities and request individuals to indicate in which ones they participated. Alternatively, an open-ended question could be asked about other types of development the individual has experienced recently. These data could be used to track which participants had participated in other development programs and analyzed to determine the contribution of participation in other development initiatives.

Specificity of Variables. Deciding exactly how to measure the specific elements of the domain identified as the area of change after a leadership development program is difficult. If a domain is measured in very specific terms that only apply to a certain group or a certain context, it will be difficult to argue that similar findings are likely in other settings. For example, many leadership development programs require participants to set goals for specific projects taking place in their organizations (for example, "hire three new members for the health care IT team by August"). Progress toward such situation-specific goals may be hard to generalize to other settings. Conversely, if the measures are couched in very general terms, they may be more applicable to different settings or groups but lack the specificity necessary to provide clear evidence for the particular group being evaluated. Generally, it is better to be certain to measure what is most important in the context of the evaluation rather than to become overly concerned about generalizing results to other situations.

Treatment Diffusion. Individuals attending a leadership development program may communicate what they are learning in the program with people outside the program. In some cases, that is an intended and valuable outcome of leadership development initiatives. For example, many leadership development programs encourage participants to share what they have learned and their development plans with their bosses in order to gain their support for the participants' change efforts. However, the situation creates a problem for evaluators because individuals who did not attend the program are experiencing elements of the program. If an individual in a unit is a member of the control group while another individual in the same unit is participating in the program, it is possible that the individual in the control group could make some improvement in his or her leadership simply because of conversations with the individual in the program. Such treatment diffusion could make the differences between the program participants and the control group appear smaller than they really are, leading to an underestimation of the program's impact.

Experimenter Effects. Conscious or unconscious actions of researchers that affect participants' performance and responses are called *experimenter effects.* An example in the context of leadership development would be an evaluator providing feedback to participants, based on observations, that improves participants' performance. While improving participants' performance is the goal of the initiative, it may be that a primary reason for the improvement was the feedback from the evaluator rather than the initiative itself. Experimenter ef-

fects can even occur simply because the evaluator discloses details of the evaluation design to participants. If participants know exactly what the evaluator is trying to measure, they may behave differently with regard to the domains being assessed.

Reactive Effects. Merely knowing that an evaluation is taking place may affect participants' behavior. These effects are sometimes called *Hawthorne effects* or *John Henry effects*. In the context of leadership development, individuals selected to participate in the program may be more confident in themselves because their organization has made an investment in them. Alternatively, if a leadership development program is perceived to be the last effort an organization makes before firing someone, participants may begin looking for other employment when they learn they have been selected to participate in a program. Increased turnover subsequent to the program may therefore not be related to the program itself, but rather the reputation of the program.

Factors to Consider in Choosing a Design

The collection of concepts and terms presented in this chapter may seem daunting, but it is not necessary for you to memorize them. Our intent is to acquaint you with a way of thinking about evaluation design that considers what kinds of questions you want your evaluation to answer and what factors might influence the evaluation's ability to answer those questions. No evaluation is perfect. This chapter can serve as a useful reference that you can refer to on an as-needed basis. You cannot successfully defend against all the threats to validity, but understanding more about the various threats enables you to better account for these elements in the design of the evaluation and in the interpretation and use of results. After all, evaluations are used to make decisions about funding and other resources, so it is important to think about the quality, accuracy, and appropriateness of the data on which decisions are based.

Following are some suggestions for incorporating the ideas presented here into the decisions you make in the design of your evaluation.

Evaluation Purpose

For an evaluation to be effective, one needs to understand the overall purpose of the evaluation from the perspective of key stakeholders. Understanding the purpose of an evaluation helps determine what kind of an evaluation is most

likely to meet that purpose. If stakeholders are looking for answers to questions related to the general questions, What changes have occurred? and Were the changes caused by the program being evaluated? then an experimental or quasi-experimental design with one or more control groups is worth considering. Keep in mind, however, that there are other methodologies that can provide information that may augment information gathered as part of an experimental or quasi-experimental design or may even be more appropriate.

Understanding what contributed to or diminished change after a leadership development initiative may be best achieved using a mixed methodology approach. Collecting stories or examples of specific changes and the barriers to and facilitators of those changes can be accomplished through interviews or focus groups. This type of information is often very helpful in providing examples that "speak" to stakeholders and can provide meaningful clues about why an initiative achieved its goals or why not. It can be even more compelling when stories collected from participants are compared to stories collected from nonparticipants for evidence of change due to the program. These data combined with quantitative data about the program can provide a more comprehensive view of the program.

Typically, quantitative data, such as 360-degree leadership ratings, are shared with stakeholders in aggregate form, indicating trends across or within groups of individuals, rather than highlighting specific individual examples or creating a deep understanding of why an initiative worked or not. Qualitative data can provide rich insight into participants' subjective experiences with the program. The evaluator's understanding of how the results of an evaluation will be used and the specific questions the evaluation seeks to answer are essential to creating or selecting the appropriate design. Both quantitative and qualitative data can be used in quasi-experimental and experimental designs.

On the following pages we present and comment on the most common experimental and quasi-experimental designs. Russ-Eft and Hoover's (2005) discussion of experimental and quasi-experimental designs provides additional information about various designs in the context of organizations.

Single-Group Designs

In a single-group design, only individuals who participated in the leadership development initiative are studied. There are essentially three ways to organize this type of design:

1. Posttest only
2. Pretest-posttest
3. Repeated measures

In the posttest-only approach, participants are measured after they have engaged in leadership development. The problem with this approach is that it is impossible to establish whether high scores are the result of the initiative or if they were preexisting. No objective measures of change can be taken, though participants can be asked to reflect and report how they think they have changed. Individuals with whom the leadership development participant interacts can also be asked to report how they think the participant changed. Keep in mind that if you are asking others about the changes in the participant, the questions need to address things someone other than the participant would be able to notice.

In contrast, a pretest-posttest approach provides information about the amount of change that occurred, although the lack of a control group still limits confidence in the program as the cause. Retrospective pretest-posttests are a variation of the general pretest-posttest approach, with the distinction being that retrospective pretests are administered after the program. In either case, it is difficult to prove the program caused the change. Any observed change might be due to another event experienced by the group, such as layoffs or annual salary increases (see the previous section on Threats to Validity for more information). If all participants show change and they are from different contexts (different sectors or organizations, for example), there may not be another plausible explanation for the change and it would therefore be easier to argue that the program caused the change.

Repeated measures designs involve collecting multiple measurements typically before the program, during the program, and after the program. Exhibit 1.3 provides an example of a longitudinal approach. Longitudinal data collected on three or more occasions allow us to track the group's scores over an extended period of time, providing evidence of trends. If there is a trend of improvement it can be more convincing than improvement measured at a single point in time. Looking at data over time is especially appropriate when there is reason to expect a dip in performance before improvement. For instance, you may expect individuals to be a bit awkward using newly acquired skills at first (they may be used to a very different style), but over time their performance might climb to a new high as they gain proficiency and comfort.

EXHIBIT 1.3. A LONGITUDINAL APPROACH TO DATA COLLECTION.

Preprogram	During the program	3 months after the program	6 months after the program	9 months after the program
Baseline measure taken	Measure progress	Measure progress	Measure progress	Measure progress

Thus, data collected at multiple points in time will reflect all parts of the performance curve.

Designs with Two or More Groups

Even when a strong repeated measures design shows change, there is still the possibility that the improvement is due to something other than the leadership development program. For this reason, evaluators should prefer designs using at least two groups.

These types of designs have the same basic variations as single-group designs (that is posttest only, pretest-posttest, or longitudinal), but at least two groups are studied in order to provide a comparison. If participants show positive change in the identified areas and those who did not participate in the initiative do not show positive change (or show less change), that provides more convincing evidence for the effectiveness of the initiative than assessing only a single group.

Ideally the groups to be compared are formed using random placement. Although it might be impractical to completely withhold a development program from some individuals at random, a random process can be used to determine the order in which individuals participate and thus achieve nearly the same result. Consider, for example, a repeating series of initiatives intended for large numbers of individuals preselected by the human resources department to participate. If the pool of selected candidates were 250, then 125 could be randomly assigned to participate as the first cohort and the remaining 125 would be assigned to participate at a later time as the second cohort. In the interim, the second group could function as the control group for the first cohort, since the groups should be similar; however, this process may not guarantee that the two groups will remain equivalent throughout the study. Members of the participant group may drop out before completing the program, or mem-

bers of either group may not be able to be located for follow-up purposes for reasons that are systematically related to the impacts of the program. Members of the control group could also be affected by treatment diffusion.

Identifying a Control Group

When random placement is not an option, there are two alternative approaches to identifying a control group: (1) matching participants and nonparticipants on important characteristics; and (2) statistically controlling for differences between groups during data analysis. Note that both of these methods require that the evaluator be able to anticipate which variables might affect the outcomes being measured. Matching participants and nonparticipants on key traits can be difficult when multiple characteristics must be considered, which is often the case with leadership development. Some characteristics to consider often include geographic location, department or function, age, gender, organizational level, and job performance indicators. If additional personal information is available, such as scores on ability, performance, or personality tests, those may also provide useful matching variables.

Often the matching can lead to other issues. For instance, if you match a participant with a nonparticipant who is in the same group and shares the same boss, you may be introducing treatment diffusion or expectation effects. The participant may talk about or model knowledge or skills gained in the initiative, thereby exposing the nonparticipant to aspects of the initiative. It is also possible that the boss may treat differently, or have different expectations of, the participant, which can lead to a change in performance (or another area measured).

Statistically controlling for differences assumes you are aware of the variables on which the groups differ and have measures to quantify them. Statistical control requires considerable statistical savvy to execute properly and likely will require additional data collection; therefore it may not be the best option in some cases. In addition, the amount of testing required to gather the data needed in the analysis may be too burdensome.

Planning for Data Analysis

Once measures have been selected or developed, and you have decided on the groups, you can begin to determine what kinds of analyses you intend to conduct. At this point, it is important to check the available sample size, especially

if generalizability is a goal. It may not be possible to obtain groups large enough to make meaningful statistical comparisons (Tourangeau, 2004). Even if the sample size was determined to be adequate prior to beginning the initiative and the evaluation, it is often the case that not all the identified individuals will complete the initiative and participate in the evaluation. In small organizations, complex statistical analyses are often not viable because it may be impossible to gather data from enough individuals to permit meaningful comparisons.

Conclusion

Used appropriately, experimental and quasi-experimental designs can be an effective tool for determining the effects of leadership development initiatives. This chapter introduces some of the core elements and issues related to using this approach as an evaluation tool for leadership development. In Exhibit 1.4, we summarize the main points of this chapter into a succinct list of recommendations for evaluators of leadership development initiatives and include a worksheet as Exhibit 1.5 to help you identify threats to internal validity that might affect your evaluation. The Resources provided at the end of the chapter are intended to help you locate additional information.

EXHIBIT 1.4. RECOMMENDATIONS FOR EVALUATORS OF LEADERSHIP DEVELOPMENT INITIATIVES.

- Use complementary methods to illustrate, qualify, and strengthen the understanding of the measured change and its causal link to the program.
- Use equivalent control groups when possible to help guard against arguments that the changes were due to something other than the leadership development initiative.
- Use multiple measures (triangulate methods and sources). If an evaluation is able to demonstrate positive impact the next question is usually "Why?" Collecting diverse information from diverse sources leads to results that are able to meet a variety of stakeholder needs.
- Make sure an appropriate sample size is available for the types of comparisons planned.

EXHIBIT 1.5. THREATS TO INTERNAL VALIDITY WORKSHEET.

Threats to internal validity are listed following, along with questions that can help you reveal and think about these threats in leadership development contexts.

Selection

- How are participants selected to participate in the leadership development initiative?
- What records are kept about the selection process?
- Is a specific type of person more likely to participate in the leadership development initiative?
- How might the selection process have an impact on evaluation efforts?

Mortality or Attrition

- What processes are in place to monitor initiative participation and track demographic (and perhaps other) information about those who complete or do not complete the initiative?
- Are resources available (for example, budget) to follow up with individuals who drop out of the initiative or the evaluation?

Statistical Regression

- If participants are selected on the basis of their performance on a measure or if there is baseline information about performance on a measure, how does the observed range of scores compare to the possible range of scores?

Maturation

- How likely is it that participants would perform better on the measures selected for the evaluation because of natural development trends (such as more time on the job)?

History

- What about the context in which individuals and groups are performing has changed or might change?
- What processes are in place to track the changes likely to have an impact on the initiative or the evaluation?
- What impact might these changes have on results?

EXHIBIT 1.5. THREATS TO INTERNAL VALIDITY WORKSHEET, Cont'd.

Testing

- Is a pretest being administered as part of the evaluation?

Instrumentation

- Is there evidence that the measures being used (for example, pretest and posttest) are reliable?
- Are the pretest and posttest the same measure? If different measures are being used, are the two measures directly comparable? If the same measure is being used, what impact might response shift have on results?

Resources

For more detailed technical information about experimental and quasi-experimental research designs, see Campbell and Stanley, 1963; Cook and Campbell, 1979; Russ-Eft and Hoover, 2005; Shadish, Cook, and Campbell, 2002.

Existing measures can be identified and evaluated using information provided by the Buros Institute of Mental Measurements, which has a searchable Test Directory online at www.unl.edu/buros/bimm. The site also provides guidance about selecting and using appropriate measures.

There are many Web sites that offer online calculators for estimating sample size requirements as well as guidance that can be helpful, for example:

www.surveysystem.com/sscalc.htm

www.stat.uiowa.edu/~rlenth/Power

www.isixsigma.com/library/content/c000709.asp

References

Campbell, D. T., and Stanley, J. C. *Experimental and Quasi-experimental Designs for Research.* Boston: Houghton Mifflin, 1963.

Collins, L. M., and Sayer, A. G. *New Methods for the Analysis of Change.* Washington, D.C.: American Psychological Association, 2001.

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

Craig, S. B., Palus, C. J., and Rogolsky, S. "Measuring Change Retrospectively: An Examination Based on Item Response Theory." In Jennifer Martineau (chair), *Measuring Behavioral Change: Methodological Considerations.* Symposium presented at the annual conference of the Society for Industrial and Organizational Psychology. New Orleans, April 2000.

Cronbach, L. J., and Furby, L. "How We Should Measure Change—or Should We?" *Psychological Bulletin,* 1970, *74,* 68–80.

Facteau, J. D., and Craig, S. B. "Are Performance Appraisal Ratings from Different Rating Sources Comparable?" *Journal of Applied Psychology,* 2001, *86,* 215–227.

Gottman, J. M. *The Analysis of Change.* Mahwah, N.J.: Lawrence Erlbaum Associates, 1995.

Harris, C. W. (ed.). *Problems in Measuring Change.* Madison: The University of Wisconsin Press, 1963.

Howard, G. S. "Response Shift Bias—A Problem in Evaluating Interventions with Pre/Post Self-Reports." *Evaluation Review,* 1980, *4*(1), 93–106.

Howard, G. S., and Dailey, P. R. "Response-Shift Bias: A Source of Contamination in Self-Report Measures." *Journal of Applied Psychology,* 1979, *64*(2), 144–150.

Leslie, J. B., and Fleenor, J. W. *Feedback to Managers: A Review and Comparison of Multi-rater Instruments for Management Development.* Greensboro, N.C.: Center for Creative Leadership, 1998.

Millsap, R. E., and Hartog, S. B. "Alpha, Beta, and Gamma Change in Evaluation Research: A Structural Equation Approach." *Journal of Applied Psychology,* 1988, *73,* 574–584.

Nunnally, J. C. *Psychometric Theory* (2nd ed.). New York: McGraw-Hill, 1978.

Rohs, F. R. "Response Shift Bias: A Problem in Evaluating Leadership Development with Self-Report Pretest-Posttest Measures." *Journal of Agricultural Education,* 1999, *40*(4), 28–37.

Rohs, F. R. "Improving the Evaluation of Leadership Programs: Control Response Shift." *Journal of Leadership Education,* 2002, *1,* 50–61.

Russ-Eft, D., and Hoover, A. L. "Experimental and Quasi-Experimental Designs." In R. A. Swanson and E. F. Holton (eds.), *Research in Organizations: Foundations and Methods of Inquiry.* San Francisco: Berrett-Koehler, 2005.

Schmitt, N. "The Use of Analysis of Covariance Structures to Assess Beta and Gamma Change." *Multivariate Behavioral Research,* 1982, *17,* 343–358.

Shadish, W. R., Cook, T. D., and Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin, 2002.

Tourangeau, A. E. *Evaluation Study of a Leadership Development Intervention for Nurses.* Final Report to the Change Foundation, Ontario, Canada, January 2004.