Evaluation Design

There are a number of models for the design and implementation of instructional programs. Most are sound and logical and produce effective learning. Some focus on impact instead of activity; others focus on the design process which emphasizes the evaluation of an instructional program, and specifically, the desired results. All models include one important step that we will explore in this lesson: selecting the evaluation method and design. This step is perhaps one of the most important among those directly involved with evaluation.

The methods of evaluation are difficult to separate from the evaluation design. However, specific methods are only a part of the process and are discussed in more detail in a later chapter. Evaluation design is a more general view of the process. It requires the development of a complete system to get the desired measurements. Proper design is imperative for effective evaluation. This brief chapter presents the basic elements of design and the various design alternatives.

Elements of Design

In many instructional program evaluations, comparisons are made in the performance of a group of program participants:

- To their performance prior to the program.
- To another group who did not receive the same program.
- To the remainder of the population of potential participants.

There can be various combinations which make up an evaluation design. This section deals with the elements of design with emphasis on the different design approaches. The specific design variations will be presented later.

Control Groups

Control groups and experimental groups frequently appear in designs. A control group is a group of participants who are as similar as possible as those in the experimental group but are not involved in the instructional program: the experimental group does participate. Ideally, the only difference between the two groups is that one participates in the program and the other does not. Therefore, a performance comparison of the two groups should indicate the impact or success of the instructional program.

It is important that the two groups be equivalent in their job settings, skills, abilities, and demographic characteristics. If possible, the identity of the control group should not be revealed. Otherwise, it could affect their performance. With these requirements, control groups may not be practical in real work situations. However, for critical evaluations affecting many participants, control groups are almost essential.

Timing of Measurements

A critical issue in evaluation is the timing of measurements (or tests). Measurements may be taken before the program, during the program, and at subsequent intervals after the program. The post-test part of a design is never omitted because it directly measures the results of a program. Careful attention needs to be given to determining when and how pretests and post-tests are conducted. In each case, participants being tested deserve an explanation about the proposed activity.
When conducting pretests, four general guidelines are recommended:

1. **Do not use the pretest if it alters the participants' performance in some way.** Pretests are intended to measure the state of the situation before the HRD program begins. The test itself should not have any effect on performance. If there is evidence that the testing procedure will affect performance, then possibly the pretest should be omitted, modified, or given far enough in advance of the program to minimize the effect.

2. **Do not pretest when it is meaningless.** In teaching completely new material or providing information that participants do not have at the present time, pretest results may be meaningless. It would just show an absence of the knowledge, skill, or ability. For example, in a foreign language program, it is meaningless to conduct a pretest in the foreign language if the participants are not expected to know it. Instead, a measurement of proficiency in acquiring the foreign language skills after the program is more appropriate (post-test only).

3. **Pretests and post-tests should be identical or approximately equivalent.** When scores are compared, they should have a common base for comparison. Identical tests may be used, although they may influence results when they are taken the second time. Similar but equivalent tests may be more appropriate.

4. **The testing should be conducted under the same or similar conditions.** The length of time and the conditions under which both tests are taken should be approximately the same.

In addition to pretesting and post-testing, tests may be conducted at different time intervals during and after a program. These tests are called time series tests and are recommended, when feasible. Time series tests after a program show the long-term effects of the program and in particular detect whether or not benefits are lasting. They are more often used in training than in education programs. Time series tests during the program measure participant progress toward the objectives.

**Threats to Validity**

When discussing the merits of various evaluation designs, it is appropriate to return to the subject of validity. There are a number of problems (or threats) which can reduce the validity of an evaluation design. Different designs can counteract or offset the effects of these threats.

**Time or history.** Time has a way of changing things. With time, performance can improve and attitudes can change — even without an instructional program. When observing output measurements of an instructional program, this question should always be asked: "Would the same results have occurred without the instruction?" This threat to validity can be isolated by modifying the evaluation design.

**Effects of testing.** It is possible that the actual experience of a test or other instrument can have an effect on performance or attitude even if no HRD program is undertaken. This effect is more likely to happen when pre-tests and post-tests are identical. With enough time between testing, participants reflect on the material and possibly even seek answers to questions that attract their curiosity. Also, after the participants know the scope of a program, they may give more favorable responses in the post-test based on their knowledge of what is expected.

**Selection.** As presented in Appendix 3, the selection of the group to participate in an instructional program can possibly have an effect on the outcome. Naturally, some individuals will perform better than others. If mostly high achievers are selected, then the results will be distorted. The same is true for underachievers. The problem can be resolved by using random selection, if it is feasible.
Mortality. Participants may drop out for various reasons. If pre- and post-tests are used, the number in the group may have changed from one testing to another. This change makes it difficult to compare the results of the two, and it is compounded by the fact that the lower-level performers are the ones who will usually drop out of a program.

These are the most common internal threats to validity. The designs presented next will attempt to overcome these threats in varying degrees.

Selecting the Appropriate Design

Selecting the appropriate evaluation design is a key part of the evaluation strategy. This section presents the common evaluation designs and the relative advantages and disadvantages of each.

One-Shot Program Design

One of the most common, and unfortunately least valid, designs involves the "one-shot" program. This technique involves a single group which is evaluated only once after an instructional program is completed. No data are collected prior to the program. There are many uncontrolled factors that might influence the measurement and invalidate conclusions based on results achieved through this design. However, the information obtained in this one-shot evaluation is better than no evaluation at all.

This design may be useful for measuring the performance of a group when there is no way to measure performance beforehand. Or, possibly, when there is no significant knowledge, skill, or ability existing before the program is conducted.

For example, a group of international sales representatives are trained to speak Spanish before conducting a market survey in Mexico. It makes little sense to evaluate their current ability to speak Spanish before the instructional program is conducted. A measurement at the end of the program to measure their ability to understand and speak Spanish is more appropriate as part of the evaluation process.

This evaluation design is useful when financial, organizational, or time constraints prohibit the use of preprogram data collection. Figure 6-1 illustrates this type of design. Since it is the least effective, it should be a minimum reference point for evaluation and used only when necessary.

Single Group, Pretest and Post-Test Design

To remedy the problem of no data to make comparisons, the next proposed design is the single group, pretest and post-test design as shown in the illustration below. This design goes one step beyond the one-step design by collecting data before and after the instructional program. The knowledge, skills, or abilities possessed by the participants before the program can be compared to the knowledge, skills, or abilities after the program to detect improvements. This is very useful when the instructor wants to find out how effective the lesson presented was. This design also can help the instructor plan the lesson content that needs emphasis when certain content is
shown as weakness of the learners in the pretest.

The effect of a pretest is one disadvantage of this design. The pretest may "tune the participants in" to the topics and questions they might not ordinarily perceive. Consequently, the changes measured by the post-test may result not only from the instructional program, but also from the fact that the pretest was taken. The effect of external factors is another disadvantage to this design. Changes in the organization, environment, the work setting, or other factors may cause changes in the performance of participants.

**Single Group, Time Series Design**

A very popular design for evaluating an instructional program involves a series of measurements, both before and after the program. This is referred to as a single group, time series design. In this design the experimental group serves as its own control group. The multiple measurements before the program eliminates some of the problems incurred when a separate control group is not used. Repeated measurements after the program not only allows for comparison of the initial results, but enables measurement of the long-term effects of the program. This design, as illustrated below, may involve any number of measurements practical for the setting. This design eliminates much of the effects of time and selection threats to validity, while the mortality threat is not counteracted.

There can be a number of outcomes in a time series design. These are illustrated below. In Outcome A, apparently the instructional program had no effect. There was no change in what was measured. Outcome B had a change in output apparently as a result of the program. Outcome C shows a brief change in output as a result of the program; however, the participants returned to the previous level of measurement. There were apparently no lasting effects from the program.
The time series design is extremely useful when measurement data are readily available as part of the organization’s performance reporting. With this design, the impact of HRD programs can be compared with previous performance over a significant period of time.