Controlling Program Evaluation

by Stephen L. Cohen, CPT, PhD

The training literature is ripe with suggestions and models for evaluating the results of programs administered in organizations. Mostly everyone in the industry can recite Kirkpatrick's four levels of training evaluation (reaction, learning, behavior, results) and Phillips' fifth level of return-on-investment. Within the last several years, however, there appears to be a heightened awareness of the need for evaluating the impact of learning interventions, caused in large part by increased visibility of the value of the human asset in the boardroom and the need to justify the economic value of training.

The reasons most organizations shy away from taking on these evaluations are too numerous to document here. Suffice it to say that relatively little formal evaluation, other than the proverbial Level 1 "smile sheets," takes place today in most typical organizations. In fact, the 2004 American Society for Training & Development State of the Industry Report indicated the following from a cross-section of benchmarking service organizations: More than 74% surveyed conduct Level 1 evaluations; 31% conduct Level 2 evaluation; 14% conduct Level 3 evaluations; and less than 10% conduct evaluation at Level 4. What appears clear, then, is that program evaluation is not easy.

Interestingly enough, while countless articles continue to appear in our trade journals, books, and presentations describing these levels and how important it is to incorporate them into training interventions, there also seems to be an absence of well-conducted research efforts demonstrating the application of these evaluation methods, particularly with the use of control groups.

To best compare the impact of training, optimal research requires the use of control groups—that is, groups that do not receive the benefit of the training intervention. Including these groups in an evaluation paradigm is an ideal way to account for the introduction of intervening and extraneous variables. True, including groups of employees in the evaluation design who do not receive training often makes it imprudent to conduct the research, since some targeted employees theoretically would not receive the training they need when they need it. And using control groups is often problematic for organizations for no other reason than the ethics of using some employees as "guinea pigs" in an experiment.

Yet many researchers consider it is nearly impossible to assess the impact of a training intervention without control groups. This article provides a brief description of the essence of experimental design and how control groups can be used in a formal training evaluation strategy.

An Experimental Design Primer

While this may go without saying for many, it is important for anyone involved in training evaluation to have at least a basic understanding of the principles of experimental design. Otherwise, faulty assumptions and conclusions may be drawn from the data, most notably that training has a decided impact on changes in behavior or business results, when in fact it does not. Evaluation methodology involves the prediction of results from the use of a particular training technique. So we must be extremely cautious, yet confident, that we are making accurate judgments about the true impact of any learning experience we put in place.

Independent Variables

At the most basic level of research design there are three sets of variables: independent, intervening, and dependent. *Independent variables* are those that can be intentionally manipulated and often represent primary factor(s) that could influence results; for example, the presence or absence of training or a certain type of training intervention. We can manipulate exactly which, if any, training intervention employees receive, just as we can change its length, approach, content, and even facilitation.

There are numerous variables that can be altered, depending on exactly what we want to evaluate. If, for example, we want to assess the impact of different learning approaches on subsequent on-the-job performance, we can compare blended versus instructor-led versus online learning approaches to assess which has the greatest impact.

Intervening Variables

Of course, we would have to control for a number of factors to be sure no one group possesses an unintentional benefit or disadvantage. These *intervening variables*, if not intentionally controlled, can alter the outcome of the research. For instance, the level of existing knowledge or skill of an employee could alter the relative impact of the type of learning approach. Highly experienced and knowledgeable employees are likely to do better than less-experienced and unknowledgeable ones regardless of the learning approach offered. So to assess which learning approach is best, one intervening variable to control would be employees' current knowledge and experience. That is, there should be a relatively similar knowledge and experience level across all the learners regardless of which approach is administered, unless we want to intentionally study the impact of these differences on the training results. Otherwise, we won't be able to isolate the different learning approach as the key factor responsible for any differences that may be achieved.

Obviously, there are myriad potential intervening variables, many of which cannot be totally controlled. In fact, we often hope that simple random assignment of people to groups, while not guaranteed, will address any differences. At a minimum, however, it is critical that many of the major variables that could have a spurious impact are proactively managed.

One of the reasons organizational research is so difficult is that it is nearly impossible to control all the social, economic, organizational, and individual differences that might surface during the focus of the research. While there are statistical methods that can help control for many of these, the extent to which they are anticipated and managed will enable drawing accurate conclusions from training impact data. Yet even if everyone goes through the same training under the same internal and external circumstances, and pre-post changes are observed in behavior, business results, or both, we really can never know if it was the training, per se, that was responsible for the changes—or some other factors that just happened to be present during the period the training was conducted. For example, major swings in the economy, pricing changes, and organizational restructuring, to name just a few, are all intervening variables. At a minimum, it appears that some people should receive training and others should not to be sure that comparisons can be made about changes in behavior between these two groups.

Dependent Variables

The third set of factors that needs to be identified is called *dependent variables*. These are the very measures we are trying to influence, whether they involve new knowledge gained, specific skill demonstration, or on-the-job performance results. It is critical that they are not only identified but stable over time; that is, they are not subject to alteration because of extraneous environmental factors. For example, some people end up performing better because they are provided with personal coaches after the training versus others who are not offered this privilege. Indeed, the training itself may not have made any difference in the behavior change, only the coaching. Other factors may influence the stability of metrics such as organizational changes, the economy, and the political environment.

With this as a backdrop, let us look at how control groups can at least begin to assist in the management of training evaluation research.

Using Control Groups

Controls refer to procedures allowing researchers to eliminate certain explanations for results. Despite resistance to using control groups to evaluate training impact, they are required to effectively assess the impact of training interventions on subsequent on-the-job performance. As noted above, there are numerous factors that can, and do, influ-

	Pretest	Training	Post-Test
Group A	Yes	Yes	Yes

Figure 1. Simple Pretest and Post-Test Evaluation Design.

ence changes in the acquisition of new knowledge, skills, and attitudes that, while having nothing to do with training, ultimately alter job performance-related behavior. These might include self-motivation to change, personal approach to learning, observation and emulation of others, natural aptitude, feedback, mentoring and coaching, and job assignments. Add to these the myriad additional factors that can affect moods, behavior, and results, such as family issues, organizational changes, immediate supervisor capability, economic conditions, changes in business strategy, market trends, regulatory issues, and so on. It is no wonder that evaluating the impact of a single training intervention on job performance has been difficult.

Typically, training evaluation consists of administering, or observing, a pretest assessment of content, behavior, or both related to the subsequent training intervention. The training is then followed by a post assessment of the effect of that same content upon behavior. This paradigm is illustrated in Figure 1. It is often assumed that if there are significant improvements in the post-test from the pretest assessments, then the training was responsible for this change.

However, as we noted above, this is not always the conclusion that can be drawn, even for statistically significant improvements, because any one of the factors noted may have been responsible—if not totally, then partially—for the increases. For example, if the training is focused on new sales skills or knowledge for an organization's sales reps and the post-test assessment of dollar growth and call volume are significantly better than the pretest scores, this change could be as much due to an upward turn in the economy as it is to the training itself. The use of *control groups* can at least statistically manage and separate the impact of many of these variables, so the unique effect of the training intervention can be assessed. A control group would comprise reps *not* subjected to the training program or process, whereas *experimental group* members would be. And, while the notion of restricting certain employees from training is appropriately anathema to any organization, there are ways to effectively evaluate the impact of training on those subjected to it, while temporarily excluding others.

Obviously, it is nearly impossible to control all variables that could influence performance after a training event. But it is incumbent upon any training community to do its very best to isolate as many of these as possible in determining if the training employed is achieving what it was designed to do.

The most common methodology for assessing outcomes, as mentioned earlier, is to ascertain whether performance, however measured, is significantly different after the training intervention versus prior to it. That is, can we deduce with reasonable certainty a change in job-related behavior between a pre- and post-training evaluation? Without a control group that is not trained, it would be impossible to statistically conclude that any pretest or post-test differences are due only to the training itself, given the myriad intervening factors that could easily have accounted for the change as well. And one such factor could even be the pretest itself. This is called *pretest sensitization* and describes the potential impact of being initially exposed to an assessment intervention that may signal expected behaviors, which could inadvertently shape subsequent performance without the benefit of the training.

Of course, one might conclude we would need hundreds of control groups to account for all the potential factors to be able to accurately assess post-training changes. But this is not necessarily the case in that we can assume in a relatively homogeneous corporate environment that many factors outside the training will, on average, randomly impact participants rather than cause severe influence on some more than others. In addition, the potential impact of other more personal factors, such as gender, age, department, tenure, and the like, can also be accounted for through statistical analyses such as analysis of variance, regression analysis, factor analysis, and so forth. Therefore, the main focus of control groups for training evaluation purposes is on the training or learning experience itself.

The methodology represented in Figure 2 is recommended for Level 2, 3, and 4 evaluations. As can be seen, in this two-group design, both groups are pre- and post-tested, but only one gets trained. This design helps isolate the training effect. That is, if there are no significant differences in the improvements from pre- to post-test assessments for the trained and untrained groups, we can safely conclude that the training itself made no difference in behavior change and is relatively ineffective in terms of achieving its intended outcomes.

	Pretest	Training	Post-Test
Group A (Experimental)	Yes	Yes	Yes
Group B (Control)	Yes	No	Yes

Figure 2. Simple Pretest and Post-Test Control Group Evaluation Design.

	Pretest	Training	Post-Test
Group A (Experimental)	Yes	Yes	Yes
Group B (Control)	Yes	No	Yes
Group C (Control)	No	Yes	Yes
Group D (Control)	No	No	Yes

Figure 3. Four Group Pretest and Post-Test Evaluation Design.

However, to draw this conclusion more reliably, there also needs to be a control for pretest sensitization—the effect of the pretest itself on the outcome of the training. As noted above, just taking the pretest could sensitize someone to what transpires, what behaviors are expected, and what outcomes are desired, all of which could interfere with the impact of the training effect. Therefore, control group participants, even though they do not receive the training, simply as a matter of being exposed to the pretest may improve their behaviors on subsequent assessments. To combat this potential effect, two more control groups, both of which would *not* be pretested, would have to be added to the design. Figure 3 illustrates this new four-group design.

Obviously, this four-group design may be difficult to execute in an organizational setting where we want everyone to be subjected to the potential benefits of the learning experience. We would never expect some people not to be trained. To manage this issue, we recommend creating a *time-sequenced* evaluation design, wherein data points are taken in time and compared as noted in Figure 4. This design lays sequencing both the training and the post-testing over the four-group design, thus achieving the control group effect, while allowing all in the target audience to benefit from the training.

In this design, the two or more training sessions (designated by 1, 2, 3, etc.) as well as the post-tests are exactly the same, just completed at a different times—days, weeks, or even months apart. This does add another dimension of control complexity in that it is possible that while unintentional, the so-called "exactly-the-same" training sessions may be more or less effective, but we need to assume that over time these differences, if any, will disappear. Furthermore, we might expect a number of intervening uncontrollable variables, not originally present, to surface between training sessions, such as organizational restructuring, new bosses, economic changes, and so on. It is also possible that those exposed to the later training sessions will do better on their second post-test simply because they had been sensitized to it in their first go-round. Again, these effects can also be managed through proper statistical analyses.

Another issue that should be addressed regarding any of these designs is there must be a sufficient number of learners to adequately apply the results to proper statistical analyses. Small classes of fewer than 10 people will make it difficult to draw any significant conclusions until perhaps at least 25 learners per group have been subjected to the chosen design.

Conclusion

In summary, it is possible to see how important it is to apply a reasonably controlled design to evaluate the impact of training whenever possible. While it appears somewhat cumbersome, the benefits of knowing whether your training is indeed achieving what it set out to achieve far outweighs the time and energy required in using control groups to conduct proper evaluations.

Stephen L. Cohen, CPT, PhD, is Vice President of the Learning Solutions Group for Carlson Marketing, the world's largest relationship marketing services firm and a division of the Carlson Companies. He is responsible for directing the design and development of Carlson's client-focused custom education and training experiences. Stephen may be reached at steve.cohen@carlson.com.

	Pretest (1)	Training (1)	Post-Test (1)	Training (2, 3, etc.)	Post-Test (2, 3, etc.)
Group A	Yes	Yes	Yes		Yes
Group B	Yes	No	Yes	Yes	Yes
Group C	No	Yes	Yes		Yes
Group D	No	No	Yes	Yes	Yes

Figure 4. Four Group Pretest and Post-Test Time Sequence Evaluation Design.