Why We Evaluate


After reading you should be able to:

- identify at least four factors affecting the success of a training program
- describe the gap between entering behavior (EB) and terminal behavior (TB)
- list at least five things to be evaluated before launching a training program
- state at least three factors to be evaluated during training
- identify at least four factors to be evaluated after training
- describe the four levels (Kirkpatrick) at which training can be evaluated
- give several examples of training you would not evaluate at Levels 3 or 4
- list at least six benefits of evaluating training.

A study released by the American Society for Training and Development (ASTD) in 1996 identified as a key issue for the new millennium the need to measure performance improvement related to training. This comes as no surprise. Top management wants to know what results the organization is getting from the hundreds of thousands of dollars spent annually in training. Instructors and course designers want to know what impact their programs are having on individuals and the organization. Trainees and their supervisors want to know what kind of payoff they can expect from taking time away from productive work to participate in a course. In short, the evaluation of training's impact is a hot topic.

A training program is most successful when the right participants (selection) receive the right knowledge, attitudes, and skills (KAS, or content) taught by means of the right methods, media, and instructor (process) at the right time (need to know) and place (location) so as to meet or exceed the organization's expectations (learning objectives and performance outcomes).

Evaluating Before You Train

Our job would be much easier if the evaluation of training's impact were a binary issue. But it's not. We're not concerned simply with whether or not we hit the target. Rather, our evaluation of impact should yield insights into such things as which course objectives need more work; where we have been successful; and where we can improve
(on design, delivery, content, length of course, selection of participants). That's why we should evaluate each of the factors listed in the preceding paragraph before a training program is launched.

Putting it another way, the evaluation process should begin with the needs analysis that precedes training. Indeed, we might define training as the process of closing the KAS gap between what our trainees bring to the course (their entering behavior, or EB) and what they must leave with to perform effectively at work (their terminal behavior, or TB). Since the purpose of training is to close this EB-TB gap, we'd better know the nature and size of this gap and the factors that will help or hinder our efforts to close it. Here, then, are some of the things we might want to evaluate before we start to design, develop, or deliver a training program:

- What do our learners want to know and expect to get from the training?
- What do our learners need to know and must get from the training?
- What competencies are required of learners? What prerequisite EB?
- What workplace factors will help or hinder the desired performance?
- What outcomes (TB) are expected? Realistic? Desirable? Measurable?
- What is the nature and size of the gap between EB and TB?
- What resources exist (people, equipment, supplies) to facilitate learning?
- What are the costs of training relative to the estimated benefits?

The answers to these and other similar questions will lead to decisions on how to make or buy the training program that will be most successful in closing the EB-TB gap. We also may be able to take actions to improve the workplace and make it more supportive of the desired TB, since many kinds of performance problems cannot be corrected by training.

**Evaluating During Your Training**

We've just examined some of the factors to be evaluated before training. Now let's look at some of the things you might want to evaluate during training so that you can take appropriate action when necessary and not wait until the course is over and it's too late:

- Are your participants comfortable? Assess seating, lighting, temperature, ventilation, breaks, pacing, mixture of theory and practice (acquisition and application, or hands-on learning).
- Are your participants learning? Use criterion tests and short quizzes to evaluate their acquisition, and practice exercises to assess their soft skills (role play, simulation) and their technical skills (at a PC or on the equipment).
• Is your content relevant? Can your participants relate the new KAS to their own needs? Can they provide examples of its practical application in their jobs? Are they active or passive throughout the course?

• Is the training enjoyable? Learning is much more effective and transfer of training from workshop to workplace is more complete when your learners enjoy the experience and contribute to its success with good participation.

Evaluating After You Train

Most participants are familiar with the end-of-course "smile sheet" that evaluates their reaction to the course. Some training programs also have a mastery test to measure how well participants have learned the material. This type of assessment is required in courses that lead to certification or licensing as a prerequisite to holding a job or performing certain kinds of tasks. Such evaluations are usually done immediately after the training.

The increased emphasis on performance in the workplace and return-on-investment (ROI), however, has led trainers to evaluate their impact months after a course is over. Only then can we get true measures of how well our graduates are dealing with the reinforcers and constraints of the workplace—the factors that help or hinder employees as they attempt to apply what they learned. Only then can we evaluate transfer of training. Here are some of the questions we might want to answer as we collect data during our evaluation of performance of our graduates at work:

• To what degree are our graduates meeting the desired TB?

• What factors are helping or hindering their performance?

• What can be done to strengthen the reinforcers and reduce the constraints?

• What aspects of our training proved to be most and least relevant?

• What changes in performance can be seen from pretraining to posttraining?

• What is the dollar value of these changes?

• How does the value of the improvements compare with the cost of training?

This list of questions reminds us that there are many factors influencing the effectiveness of training that can best be measured after some time has elapsed following the training program. But how much time? Are we talking about weeks, months, or years? The answer is critical if we plan to do an accurate cost-benefit analysis to measure the ROI. And here we have both good news and bad.

The good news is that the more times a training program is run, the greater will be the number of graduates. The costs of running a course are usually known in advance, before the first cycle is launched: conducting the needs analysis, selecting or creating the course, running the field tests, making revisions, reproducing the materials. These are
fixed, one-time costs. In contrast, the benefits are variable as a function of (a) how many people are trained over the life of the course, and (b) how long each graduate will be able to use the new learning on the job (the pay-back period). This usually means that by waiting until a course has been conducted many times and run through its life cycle, we will have trained the largest number of employees who can yield a return on the initial investment. We will have increased our chances of getting the best ROI possible.

The bad news is that the longer we wait to evaluate performance in the workplace attributable to training, the larger will be the influence of intervening variables—factors that have nothing to do with the quality of our training but that affect performance and make it impossible for us to tell what behavior is and isn't attributable to training. Some examples are changes in technology; management; systems and procedures; organizational priorities and objectives; and the composition of the workforce.

In courses where innovation and technology lie at the core of the training, the life expectancy and payback period of a course may be measured in months. But in courses dealing with skills and competencies that should last a lifetime, the payback period will cover years. Such courses include leadership, team building, project management, writing skills, interpersonal communications, problem solving, negotiation, and conflict management. Isn't it ironic that many of the so-called soft skills courses have a longer payback period than many of the high-tech courses that are very situational, specific, and often "here today, gone tomorrow."

### Evaluating at Four Levels

We've just examined the things we might want to evaluate before, during, and after running a training program. We've looked at many questions to be answered at each of these three phases of human resources development (HRD). Now let's turn our attention to the four levels at which training can be evaluated. And here we are indebted to Donald Kirkpatrick of the University of Wisconsin for the model he developed almost 40 years ago. A complete explanation can be found in his book, *Evaluating Training Programs* (San Francisco: Berrett-Koehler, 1994). Here are the four levels of the Kirkpatrick model:

<table>
<thead>
<tr>
<th>Level</th>
<th>Issue</th>
<th>Questions Answered</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reaction</td>
<td>How well did they like the course?</td>
<td>Rating Sheets</td>
</tr>
<tr>
<td>2</td>
<td>Learning</td>
<td>How much did they learn?</td>
<td>Tests, Simulations</td>
</tr>
<tr>
<td>3</td>
<td>Behavior</td>
<td>How well did they apply it?</td>
<td>Performance Measures</td>
</tr>
<tr>
<td>4</td>
<td>Results</td>
<td>What return did the learning investment yield?</td>
<td>Cost-Benefit Analysis</td>
</tr>
</tbody>
</table>

Here's another way of depicting the four-level model. As you study it think about several of the training programs with which you are familiar. For each one, recall the levels at which the impact is and isn't being evaluated. Then return to the observations noted below.
Data is easiest to generate at the top of the model and progressively harder to collect and to interpret (that is, attribute to training) as we go down the model. The effects being measured are short range at the top of the model and progressively longer range as we descend (that is, more permanent and more delayed before we can observe and measure them). Estimates indicate that over 85% of all training programs are evaluated at Level 1. This number drops progressively as we descend, with fewer than 10% being measured at Level 4.

These observations should not surprise us. Levels 1 and 2 are concerned with the learner's behavior in class or upon completion of the training. Thus, they can usually be measured by relatively simple paper-and-pencil exercises: end-of-course evaluation sheets for Level 1, and mastery tests for Level 2. Such instruments are relatively easy to create, to administer, and to interpret (tally, score, summarize).

In contrast, Levels 3 and 4 are concerned with the learner's behavior after the training is over—sometimes long after, as in courses on Selection Interviewing or Performance Appraisal or Selling to a Multinational Client, where it may be months or years before the trainee has a chance to apply what was learned in class. Although trainers attempt to apply the just-in-time concept and want to offer their programs at a time and place that enables learners to take courses just before they need them, this is virtually impossible to do in most organizations. Thus, any failure to get satisfactory results when we evaluate at Levels 3 and 4 may be attributable to the decay due to time lapse (a natural consequence of all training and education) and not to any inadequacy on the part of the learner, the instructor, or the course design.

Self-instructional courses via text or computer or interactive video (computer-based training, touch screen or "info-window") can help to overcome the delay factor by making training available when the learner needs it rather than when the instructor and organization offer it. But self-study has its limits with learners who need the personal touch and with courses that require hands-on learning or interaction with other learners (role plays, case method, games or simulation, PC operations, lab work).

**Workshop vs. Workplace**

The instructor and course designer have a pure environment in class or self-study to take readings at Levels 1 and 2. The learners are captive, relatively free of workplace distractions, and usually motivated to learn— or at least to keep the instructor happy enough to get through the training without hassle or incident.

In contrast, their performance back in the workplace (as measured by Levels 3 and 4) is influenced by many variables not addressed in class. They operate in a contaminated environment.
environment, and no one can control the intervening variables (or even hold them constant or equate for them) so that we might attribute the learner's performance to training alone. Rather, behavior on the job must be attributed to the whole array of factors influencing how people perform at work. Here are 10 of these factors, although the list is by no means exhaustive:

1. relevancy of what was taught to what is needed to perform
2. immediacy of opportunity to apply (discussed earlier)
3. supportiveness of immediate supervisor of the learner
4. degree to which peers practice what the learner is trying to apply
5. time, money, and resources to support the new behavior
6. rewards and punishments that reinforce behavior
7. agreement of equipment and procedures between workshop and workplace
8. culture focused on long-range performance
9. workplace free of distractions, interruptions, and physical constraints
10. immediacy and specificity of feedback.

When To Measure at Levels 3 and 4

Although the list above might be enough to discourage the staunchest trainer from going beyond the use of smile sheets and end-of-course tests, there are a number of situations in which the behaviors elicited in class are virtually identical to the performance required on the job. In other words, environmental contamination is minimal. In such cases, Level 3 and 4 data may be relatively easy to obtain. Courses that deal with the following are examples:

- safety/drugs/alcohol (number of occurrences before and after training)
- data processing (order entry, inventory control)
- assembly line procedures (soldering, crimping, screwing, gluing)
- processing of paperwork (bank tellers, insurance claims processors)
- customer service (representatives behind desk in banks, airports, hotels)

These jobs are ones where 'the rubber hits the road," where it is possible to measure the quantity and quality of output-number of transactions per hour, number of errors (rejects, overages or shortages, reworks) per 1,000 units or transactions, and so on.
Let's look at another type of job where Level 3 and 4 measurement might be feasible: persons who have a high degree of control over how their time is spent, and whose work is relatively independent of direct supervision or the intervention of other employees. Again, here are some examples:

- salespersons on commission (real estate, insurance, manufacturer's representatives)
- trades and crafts (electrician, plumber, carpenter)
- creative work (computer programmers, designers, writers, consultants)
- truck delivery (route drivers, postal service, messengers or couriers)

As we look over the types of jobs in which the 10 factors listed earlier are present or absent, we come to several conclusions. The following serve as general guidelines in deciding whether or not to measure at Levels 3 and 4:

1. The farther down the organization chart you are training, the easier it is to collect data on performance at work and its impact on the organization's mission and bottom line. Conversely, the farther up you go supervisory, managerial, executive), the harder it is to quantify and to observe output.

2. The more directly responsible employees are for their own output and the more influence they have over the variables affecting output (line jobs vs. staff jobs, producing goods vs. services), the easier it is to go to Levels 3 and 4.

3. The more control employees have over how they spend their time (service repair, taxi driver, consultant), the more accountability they have for demonstrating results and ROI at Level 4.

The more direct influence an employee's performance has on earnings (commissions, bonus, incentives), the greater the likelihood that the employee will work extremely hard at overcoming workplace factors- our list of 10-that get in the way of outstanding performance (sales commissions, stock options, profit sharing).

**When To Skip Levels 3 and 4**

To the results-oriented manager or owner who is attracted to Adam Smith's economics, there's a certain appeal to the argument that Levels 1 and 2 don't count: "We're not paying employees to like the course or, for that matter, to learn. We're paying them to perform, and training has added no value to the organization until it can demonstrate that it produced improved performance in the workplace and contributed to the attainment of our mission and objective a return on the investment."

There are several fallacies to this argument. They are identified by the four "facts" that follow. When these facts are present in your training programs, it may mean that you should either skip data collection in the workplace or settle for opinions of impact (subjective and soft data) rather than measures of performance (objective and hard data).
Fact A: Some workplace behaviors cannot be measured objectively. We teach supervisors how to do performance appraisals that are constructive, interactive, supportive, focused on performance rather than personality, and legal. But how can the course's impact be measured? We cannot observe our graduates as they conduct appraisals; it's a rather private affair. Even if both parties agreed willingly to be observed, the presence of an observer would contaminate the appraisal, and we would not get a reliable reading. True, we might examine the comments entered on the appraisal form after the review, assuming we get permission. This might show us whether our supervisors have focused on performance rather than on personality. But we are still left without any objective measure of the review itself. So we can try for subjective, soft data (by surveying supervisors and their employees, getting their opinions on pretraining and posttraining behavior during reviews). Or we can abandon attempts to measure at Levels 3 and 4.

Fact B: Some workplace behaviors may never occur. Examples include how to give artificial respiration; how to operate a fire extinguisher; how to use karate or jujitsu on an assailant; how to handle a bank holdup; how to terminate an employee; how to evacuate a 727 airplane in a water landing; how to tow a drowning person to shore; how to manage a nuclear power plant reactor in a crisis. The ultimate example, of course, is military training. Governments spend billions of dollars annually training the armed services to perform complex operations under combat conditions that, God willing, they will never have to do. In all of these situations, the usual way to assess at Level 3 is to create simulations -- a sort of fail-safe approximation of the real thing. But the trainees know it isn't real, and the emotional response is very different.

Fact C: Some workplace behaviors cost too much to measure. We teach supervisors how to do selection interviews. We could follow up by sending a professional evaluator (consultant or psychologist) as a job applicant to each trained supervisor who had a job opening during the three months following training (along with actual applicants, of course, that the personnel department would screen and provide). But this would be expensive, time consuming, and lead to questions of ethics and appropriateness. What if supervisors find out that they are being "shopped," for example? What effect will this have on their level of trust in training and HRD? Another example is the use of assessment labs before and six months after supervisory training. While this might be an excellent way to measure impact, the cost would be prohibitive.

Fact D: Most workplace behaviors cannot be attributed to training alone. In our earlier list of factors influencing performance, we made this point. When behavior is significantly better following a course, trainers are quick to take credit. When it isn't, we cite all the workplace factors that intervened (much like the advertising agency that claims credit for an ad campaign when sales go up, but cites recessionary trends and a flat economy when sales fail to result). The pie diagram above reminds us that the training employees receive is but one factor affecting their performance. The size of each wedge differs with each course and each job. The larger the influence of other wedges relative to the size of the training wedge, the harder (and more dangerous) it is for trainers to attribute workplace performance to workshop efforts.