How to Evaluate Training

Using the Kirkpatrick Model

By Phil La Duke

A worker’s core competency may be the best predictor of safety. In many cases, training is slapped together and poorly delivered. But, how can one accurately assess the efficiency of training? The Kirkpatrick model is the worldwide standard for evaluating the effectiveness of training” (Kirkpatrick Partners, 2017). “It considers the value of any type of training, formal or informal, across four levels. Level 1, Reaction, evaluates how participants respond to the training. Level 2, Learning, measures if they actually learned the material” (Kirkpatrick Partners, 2017).

The Kirkpatrick model is a simple and fairly accurate way to measure the effectiveness of adult learning events (i.e., training), and while other methods are introduced periodically, the Kirkpatrick model endures because of its simplicity. The model consists of four levels, each designed to measure a specific element of the training.

**Level 1: Reaction**

Kirkpatrick’s first level measures the learners’ reaction to the training. A strong correlation exists between learning retention and how much the learners enjoyed the time spent and found it valuable. Level 1 evaluations are typically completed immediately at the conclusion of the course using what trainers euphemistically call a “smile sheet” (references how many smiles you counted at the end of a class). But, a good Level 1 evaluation should delve deeper than merely whether people liked the course. A good course evaluation will concentrate on three elements: course content, the physical environment and the instructor’s presentation/skills.

One can gain important insights into course quality if the course evaluation is effective. Typically, this means using a Likert scale (asking participants to match their agreement with a statement about the course using a scale of 1 to 5, where 1 indicates strong disagreement and 5 strong agreement). To build an effective Level 1 tool, always have statements that are positive so that a score of 1 is consistently bad and a score of 5 is consistently good. Write the statements in complete sentences and do not ask questions. Also, do not write more than 10 statements, as people tend to want to get out of the class as quickly as possible. Include two questions to end the evaluation: 1) What did you like most about the course; and 2) What could be improved?

**Level 2: Learning**

Level 2 of Kirkpatrick’s model is learning, that is, how much of the content attendees learned as a result of the training session. This evaluation is typically achieved through the use of a pre- and posttest. This causes consternation among people who do not understand how to evaluate training. Many organizations refuse to test workers and even those that do may balk at the idea of a pretest.

Pre- and posttests are key to ascertaining whether the participants learned anything in the learning event. Identical pre- and posttest scores indicate the amount of learning that took place. Without a pretest, one does not know if the trainees knew the material before the session, and unless the questions are the same, one cannot be certain that trainees learned the material in the session. It is important to shuffle the questions and also display the answer choices in a different order to prevent people from memorizing the choices without thinking about the information.

I prefer a 20-question, multiple-choice pre- and posttest because the chance of guessing a single true/false question is 50%, but that assumes the question does not contain any language that tips people off.

“True or False: It is never safe to work on energized equipment without locking out.” It is easy to guess true/false questions that include absolutes such as never, because for the statement to be true no possible scenario must exist in which the statement is true. In other words, if I can find one case where it is safe to work on energized equipment (e.g., during test mode or other conditions that require power), I can be confident that the statement is false. Conversely, we need to have a clear definition of safe; if by safe we mean the absolute absence of risk of injury (a circumstance that is nearly impossible) we can be confident that the statement is true.

Well-written multiple-choice questions provide a clearer picture of whether trainees learned. For example, a pretest question might read:

- The element with the lowest atomic weight is:
  a) Hydrogen
  b) Argon
  c) Helium
  d) I don’t know

Some may laugh at using “I don’t know” as a distractor (wrong choice), but people select that answer surprisingly often. Certain elements make this a good question: there is only one correct answer, and the distractors are correct answers to other questions. Here, again, proper grammar makes a difference. If a question read like this:

- An element whose oxidation number is 0 prevents gas from forming compounds readily, is called an ________ gas:
  a) inert
  b) low reaction
  c) nonreactive
  d) I don’t know

The Kirkpatrick model endures because of its simplicity.
Since the word *an* is used directly before the blank, basic grammar tells us that the correct answer begins with a vowel. Using common sense, one can assume that d) I don’t know is incorrect. By process of elimination, we can conclude that a) inert must be the correct answer, as all other possibilities are grammatically incorrect. Never use answers such as “a) and b) only” or “all of/none of the above” because you risk testing reading comprehension skills instead of knowledge acquisition.

Some critics think 20 questions is too many or that they take too long to answer. But, 20 questions is enough to be statistically valid (assuming a couple of variables). Let’s assume 20 people are in the class and each person is taking a 20-question test. For a confidence level of 95% and a confidence interval of ± 5, one would need a population of 19. Once the test is validated, one can be certain that the difference between the pre- and posttests is the result of the learning event.

When analyzing the test scores, understand they should be skewed to the right (i.e., the test scores should be disproportionately high, indicating that most people mastered the content). One can further analyze the data using \( \mu \) scores. A \( \mu \) (pronounced moo) score is the average of the averages; scores that are going up indicate that the instructors are getting better at their jobs. Scores that are going down indicate instructors are getting bored, taking shortcuts or for some other reason failing to present the full content. Assuming the content and the test have not changed, the \( \mu \) score is an accurate reflection of the instructor’s performance.

The first two levels of Kirkpatrick’s model are how OSH trainers can validate whether their training is effective.

**Level 3: Behavior & Level 4: Results**

Levels 3 (behavior) and 4 (results) of Kirkpatrick’s model are a bit complex and in-depth discussion is of limited value since so few get the first two levels correct. Level 3 measures whether the learning is transferred into practice in the workplace. This kind of measurement can be tricky, yet many people skip the first two steps in favor of jumping to the third. Not only is this shortsighted, it might also be dangerous.

It is important to measure the first two levels. Once it is assured that the learners understand the purpose and importance of the class they have completed, and it is known whether they acquired skills as a result of the course, one must determine whether they apply those skills in the workplace.

Correct application of skills is especially important in safety because it can literally make the difference between life and death. All training should be evaluated, and the concept of pre- and posttesting can be confusing to some. At times, for safety’s sake, trainers must assume the worker knows nothing when they evaluate whether non-classroom training is effective. The cost-effective way to deliver this training is through demonstration and practice. In this type of training, a skilled trainer or veteran employee uses a task list or a standard work instruction sheet to demonstrate the correct way to perform the task while the learner watches and asks questions until s/he is ready to practice the tasks under the watchful eye of the trainer.

In cases where the worker could injure him/herself by operating heavy equipment or unfamiliar machinery, administering a pretest that allows the learner to try to operate a machine or induction molder would be reckless and provide no value except to determine whether the worker already has the skills that the shadow training is meant to impart. Trainers must err on the side of caution (not to mention avoiding legal liability and violation of the General Duty Clause) and provide the training to the person.

Some trainers and especially veteran employees tend to skip areas of training or skim on demonstration and practice because the worker seems particularly adept at the skills. This makes the Level 3 evaluations critical. I tend to wait a month, then ask the worker to perform those tasks on which s/he was trained to learn whether the worker is still able to perform the skills to standard. During this evaluation, I ask specific questions that align with the course objectives to determine what has been retained.

At this point, some may wonder whether it is necessary that the worker retain the information presented in an orientation. The problem with an orientation is that learners frequently lack sufficient context to synthesize the concepts presented. This may sound like psychobabble—after all, what sort of context does one need to understand that fire can burn you and electricity can shock you?

Unfortunately, after a month on the job a worker often starts to feel invincible and the once-terrifying becomes mundane or even acceptable. By conducting a Level 3 evaluation, one can both evaluate training effectiveness and provide context and much-needed reinforcement of critical safety points that the training was intended to impart. With repeated evaluations and reinforcement, the training can ultimately become internalized and hardcoded into the behavior.

There is a complex connection between evaluation and reinforcement. When trainers get to Level 3 evaluation, this intricate connection is especially strong. As for Level 4 (results), let’s leave that to the training professionals to debate.

**References**