How To Start An Objective Evaluation Of Your Training Program

DONALD L. KIRKPATRICK, Ph.D.
Assistant Professor, Industrial Management Institute
University of Wisconsin

Most training men agree that it is important to evaluate training programs. They also feel that the evaluation should be done by objective means. However, the typical training man uses evaluation sheets or comment sheets as the sole measure of the effectiveness of his programs. He realizes he should do more, but he just doesn't know how to begin an objective evaluation.

According to Raymond Katzell, a well known authority in this field, the evaluation of a training program falls into a hierarchy of steps that can be briefly stated as follows:

Step One: To determine how the trainees feel about the program.

Step Two: To determine how much the trainees learn in the form of increased knowledge and understanding.

Step Three: To measure the changes in the on-the-job behavior of the trainees.

Step Four: To determine the effects of these behavioral changes on objective criteria such as production, turnover, absenteeism, and waste.

In climbing this ladder of evaluation, most trainers have completed the first step. Typically, the training director asks the trainees to fill out evaluation sheets at the end of the program. Questions that are asked most frequently are:

1. How do you rate the program?
   .....Excellent .....Very Good
   .....Good .....Fair .....Poor
2. What subject did you like best?
3. What subject did you like least?
4. What did you learn that you can use on the job?
5. What subjects would you like to have discussed at future programs?

Usually the trainees are not asked to sign their name for fear they will not give an honest reaction.

This kind of subjective evaluation is important. It gives a good indication of how the trainees reacted to the program. If they react favorably, the trainer can justifiably pat himself on the back and say, "I gave them a program they liked." But he can't rightfully claim that the training program accomplished the objective, unless his objective was to give them a program they liked.

The immediate objective of any training course can be stated in terms of the

desired knowledge and understanding that the program is trying to impart to the trainees. It is this stage of evaluation that should be undertaken as the second step. It is much more difficult than step one and, therefore, is not undertaken by many trainers.

Among the possible methods for determining whether increased knowledge and understanding have taken place, the best one seems to be the "before and after" paper and pencil test. If the scores on the posttest are significantly higher than on the pretest, the course can be deemed effective.

IN determining the effectiveness of the training, it is important to note that the paper and pencil test or inventory must cover the principles and facts that are discussed in the course. If the trainer can find a test that covers this material, he can use it. If he cannot find a suitable one, he must construct his own inventory. Some of the inventories that are available are: *How Supervise?* by File and Remmers; *Supervisory Inventory* by Wesley Osterberg; and the *Supervisory Inventory On Human Relations* constructed by this writer.

So far, then, it has been stated that a before and after test can be used to determine whether or not increased knowledge and understanding have taken place. Also, that the inventory should cover the course content. In order to determine whether or not an available test is suitable, a trainer must examine his course outline and list the principles and facts he is trying to teach. A comparison of test items with these objectives will reveal whether or not the test can be used. Because the construction of a test involves such factors as the choice of items, the wording of items, the number and type of possible response, and the sequence of items, it is far better to use an available inventory if it covers most of the course content.

Having selected or constructed a test, the trainer should consider some "Do's" for administering it:

1. Give the pretest at the start of the first class and the posttest at the close of the last session. This will minimize the influence of factors apart from the training course.

2. Have the trainee sign both the pretest and posttest. Then, the increased knowledge and understanding can be computed for each individual.

3. In instructing the trainees before they take the pretest:
   a. Tell them it is a before and after procedure.
   b. Explain the purpose of the test.
   c. Encourage them to answer truthfully by assuring them that their scores will have no effect on their pay or status in the company.
   d. Tell them to answer every question even if they have to guess. (This will be taken into account in the statistical analysis of scores.)
   e. Encourage them to take their time in taking the test. This will help to motivate them to read each item carefully.

In analyzing the test results, there are two kinds of evaluations to be made:

1. Was the entire course effective as shown by gains from pretest to posttest scores for all trainees?

2. What specific facts and principles were learned as shown by changes from pretest to posttest for each item?

Overall Effectiveness of the Course

In considering the first question, the total score of correct responses for each individual is determined for his pretest and posttest. These scores are compared to determine his gain from pretest to posttest. The mean gain (Mg) is then computed as illustrated in Table 1.

TABLE 1
Computation of Mean Gain

Trainee	Pretest Score	Posttest Score	Gain
1. J. Anderson	79	85	6
2. W. Brown	84	93	9
3. K. Dalberg	80	94	14
4. M. Fulton	68	79	11
5. G. Gage	94	95	1
6. V. Grenfell	86	88	2
7. B. Howard	79	90	11
8. J. Lewis	85	91	6
9. R. Mason	85	87	2
10. S. Stanley	60	78	18
Mean	80	88	8

The next step in determining whether or not the changes in scores from pretest to posttest are significant enough to prove the program effective is to calculate the Standard Deviation (S.D.) of the Mg. This can be done by using the following formula:

$$S.D. = 1/N \sqrt{NEX^2 - (EX)^2}$$

where:

N = the number of individuals being measured
X = the gains
EX^2 = the sum of the squares of the gains
$(EX)^2$ = the square of the sums of the gains

Using the figures from TABLE 1 we find the following:

X	X^2
6	36
9	81
14	196
11	121
1	1
2	4
11	121
6	36
2	4
18	324
80	924

$N = 10$
$EX^2 = 924$
$(EX)^2 = (80)^2 = 6400$

$S.D. = 1/10 \sqrt{10 \times 924 - 6400}$

$S.D. = 5.3$

The next step is the computation of the **estimated standard error of the mean** (Sm) by the following formula:

$$Sm = \frac{S.D.}{\sqrt{N-1}} \qquad Sm = \frac{5.3}{\sqrt{10-1}} = 1.8$$

The "t" score can then be determined by the formula:

$$t = \frac{Mg}{Sm} \qquad t = \frac{8.0}{1.8} = 4.4$$

Reference to a "t" table such as can be found in McNemar's *Psychological Statistics* reveals that this "t" of 4.4 corresponds to a probability (P) of less than .01. This means that for our example described in TABLE I, the mean gain of 8 is so large that it can be attributed to chance less than one time in one hundred. According to accepted practice, if the P is .05 or less, the gain is significant and the course can be termed effective. It follows that the less the P, the more effective the course. Therefore, the example given in this article proves that the course was very effective.

Facts and Principles That Were Learned

Of equal importance in evaluating the course is the determination of which specific facts and principles were learned by the trainees. In other words, were there a few facts that were learned by most of the trainees or did different trainees learn different things? This kind of evaluation will reveal the effectiveness of the instructors in getting across specific points.

In order to measure this, the pretest scores of each item must be compared to the posttest scores for that item. The significance of change can be determined by the relatively simple chi square formula:*

*This formula applies where there are only two possible responses to each item, an "Agree" and a "Disagree." The writer's "Supervisory Inventory on Human Relations" is an example of this kind of a test.

$$X^2 = \frac{(A-D)^2}{A+D}$$

in which:

X^2 = chi square
A = changes from "Agree" on pretest to "Disagree" on posttest
D = changes from "Disagree" on pretest to "Agree" on posttest

In case A + D totals less than 10, the following formula should be used:

$$X^2 = \frac{[(A-D)-1]^2}{A+D}$$

Reference to a chi square table that also can be found in McNemar reveals the probability.

As an example, Item 1 on the inventory states:
"The best way to train a new worker is to have him watch a good man on the job."

A tabulation of responses to this item reveals that 10 individuals changed their responses from "Agree" on the pretest to "Disagree" on the posttest. One individual changed his response from "Disagree" on the pretest to "Agree" on the posttest. The rest of the individuals responded with the same answer on both tests. The question is, did the responses on this item change significantly enough to prove that the instructor got across his point?

Substitution in the formula reveals:

$$X^2 = \frac{(A-D)^2}{A+D} = \frac{(10-1)^2}{10+1} = 7.4$$

Reference to the chi square table shows that this change has a probability of less than .01. Therefore, the trainer was successful in teaching the principle that there is a better way of training a new worker than to have him watch a good worker on the job.

A similar analysis of each item will reveal which facts and principles were learned by the trainees.

Summary

Training men agree that it is advisable to evaluate training courses as objectively as possible. Typically, their evaluation consists of subjective comment sheets that are completed by trainees at the end of the course. Providing that these are properly administered, these evaluation sheets give a valid measure of trainee reaction to the program. However, they do not give any evidence of benefits derived.

The first step in objectively evaluating the effectiveness of a training course is to determine whether or not the desired facts and principles were learned by the trainees. This can be done by:

1. Using a suitable paper and pencil test.
2. Testing the trainees before and after the program.
3. Determining the overall effectiveness of the course by comparing pretest and posttest scores for each trainee.
4. Determining which specific facts and principles were learned by analyzing the changes on each test item from pretest to posttest.

The purpose of this article is to suggest a specific technique for beginning an objective evaluation of a training program. Further efforts should be undertaken by every training man to follow up this kind of an evaluation by attempting to measure trainee change in behavior that occurs as a result of participation in the program.
