FOCUSING ON REACTION, LEARNING, BEHAVIOR AND RESULTS . . .

TECHNIQUES FOR EVALUATING TRAINING PROGRAMS

BY DONALD L. KIRKPATRICK
(JUNE, 1960)

This series of articles is based on the following assumption: That *one training director cannot borrow evaluation results* from another; *he/she can,* however, *borrow* evaluation *techniques.* Therefore, the techniques used by various trainers will be described without detailing the findings. Each of these four articles will discuss one of the evaluation steps which can be summarized as follows:

Step 1 — REACTION
Step 2 — LEARNING
Step 3 — BEHAVIOR
Step 4 — RESULTS

These articles are designed to stimulate training directors to increase their efforts in evaluating training programs. It is hoped that the specific suggestions will prove helpful in these evaluation attempts.

The following quotation from Daniel M. Goodacre III is most appropriate as an introduction:

"Managers, needless to say, expect their manufacturing and sales departments to yield a good return and will go to great lengths to find out whether they have done so. When it comes to training, however, they may expect the return — but rarely do they make a like effort to measure the actual results. Fortunately for those in charge of training programs, this philanthropic attitude has come to be taken for granted. There is certainly no guarantee, however, that it will continue, and training directors might be well-advised to take the initiative and evaluate their programs before the day of reckoning arrives."

PART 1 — REACTION

Reaction may best be defined as how well the trainees liked a particular training program. Evaluating in terms of reaction is the same as measuring the feelings of the conferees. It is important to emphasize that it does not include a measurement of any learning that takes place. Because reaction is so easy to measure, nearly all training directors do it.

However, in this writer's opinion many of these attempts do not meet the standards listed below:

1. Determine what you want to find out.
2. Use a written comment sheet covering those items determined in step one above.
3. Design the form so that the reactions can be tabulated and quantified.
4. Obtain honest reactions by making the forms anonymous.
5. Encourage the conferees to write in additional comments not covered by the questions that were designed to be tabulated and quantified.

The comment sheet shown in Figure 1 was used to measure reaction at an ASTD Summer Institute that was planned and coordinated by the staff of the Management Institute of the University of Wisconsin.

Those who planned this ASTD program were interested in reactions to: subject, technique (lecture vs. discussion), and the performance of the conference leader. Therefore, the form was designed accordingly. So the reactions could be readily tabulated and quantified, the conferees were asked to place a check in the appropriate spaces.

In question 3 concerning the leader, it was felt that a more meaningful rating would be given the leader if the conferees considered items A through G before checking the "overall rating." This question was designed to prevent a conference leader's personality from dominating group reaction.

Question 4 allowed the conferees to suggest any improvements that came to mind. The optional signature was used so that follow-up discussions with conferees could be done. About half of the conferees signed their names. With this type of group, the optional signature did not affect the honesty of their answers, in all probability. It is strongly suggested that unsigned sheets be used in most meetings, however.

This ASTD reaction sheet was used at the conclusion of every session in the institute program. Therefore, the conferees rated each conference leader for his contribution to the program. In many internal training programs, a series of meetings will be held and the reaction sheet will not be used until the end of the last session. This is especially true when one conference leader conducts the entire program. In this case, a comment sheet like the ASTD one might be adapted to the situation and modifications made.

How to Supplement the Evaluation of the Conferees

In doing research on the subject of evaluation, this writer received a very practical suggestion from Richard Johnson, past president of the New York chapter of ASTD. Mr. Johnson suggested that the comment sheets be given to the enrollees before the program is over so that the suggestions can be used in improving the last section of the training program. For example, where a training program consists of a series of nine sessions, the comment sheet should be given to conferees at the end of the third session. Their comments and suggestions should be taken into consideration to make the last six sessions more effective.

So far in this article, the techniques for measuring the reactions of the enrollees have been discussed. It has been emphasized that the form should be designed so that tabulations can be readily made. In this writer's opinion, too many comment sheets are still being used in which the conferees are asked to write in their answers to questions. Using a form of this kind, it becomes very difficult to summarize comments and to determine patterns of reaction.

At The Management Institute of the University of Wisconsin, every session is evaluated in terms of the reactions of the conferees. This has been done for more than 10 years. Many times, the coordinator of the program felt that the group reaction was not a fair evaluation of the effectiveness of the program. Sometimes the staff men felt that the conference leader's personality made such an impression on the group that he received a very high rating. In other sessions, the coordinator felt that the conference leader received a low rating because he did not have a dynamic personality. Frequently, in the opinion of the coordinator, the

Figure 1

RATING CHART

Leader____________________ Subject____________________

Date________________

1. Was the Subject Pertinent to Your Needs and Interests?

☐ No ☐ To Some Extent ☐ Very Much So

2. How Was the Ratio of Lecture to Discussion?

☐ Too Much Lecture ☐ O.K. ☐ Too Much Discussion

3. How About the Leader?

	Excellent	Very Good	Good	Fair	Poor
A. How well did he/she state objectives?					
B. How well did he/she keep the session alive and interesting?					
C. How well did he/she use the blackboard, charts, and other aids?					
D. How well did he/she summarize during the session?					
E. How well did he/she maintain a friendly and helpful manner?					
F. How well did he/she illustrate and clarify the points?					
G. How was his/her summary at the close of the session?					

What Is Your Overall Rating of the Leader?

☐ Excellent ☐ Very Good ☐ Good ☐ Fair ☐ Poor

4. What Would Have Made the Session More Effective?

____

____

____

____

Signature____________________________________

Figure 2

LEADER RATING SHEET

Rating______ Date________________ Rater's Initial______________

Name of Leader________________________ Subject______________

	Very Much So	To Some Extent	No
A. PREPARATION			
1. Did he/she prepare for the meeting?			
2. Was his/her preparation geared to the group?			
B. CONDUCTING			
1. Did he/she read the material?			
2. Did he/she hold the interest of the group?			
3. Was he/she enthusiastic/dynamic?			
4. Did he/she use visual aids? If yes, what aids?			
5. Did he/she present the material clearly?			
6. Did he/she help the group apply the material?			
7. Did he/she adequately cover the subject?			
8. Did he/she summarize during conference and at end?			
9. Did he/she involve the group? If yes, how?			

C. CONSTRUCTIVE COMMENTS

1. What would you suggest to improve future sessions?

____

____

D. POTENTIAL

1. With proper coaching what would be the highest rating he/she could achieve?__________

E. ADDITIONAL COMMENTS

____

____

Figure 3

OSCAR MAYER & CO. EVALUATION FORM

Key	*Programs*
A	Modern Leadership for Middle Management
B	Supervisors' Leadership in Cost Control
C	Developing Supervisory Skills
D	Human Relations for Foremen & Supervisors
E	Leadership and Growth
F	Creative Thinking for Supervisors
G	Human Relations for New Foremen
T	Totals

	A	B	C	D	E	F	G	T
Questionnaires returned:	3	3	5	11	5	1	1	29
1. I thought the program was:								
A. Very well organized and helpful	3	3	5	11	5	1	1	29
B. It was of some value								
C. It was poorly organized and a waste of time								
2. In reference to the subject content:								
A. It was all theory and of little practical value								
B. It was both theory and practical	3	2	2	3	1			11
C. It was very practical and useful	0	1	3	9	4	1	1	19
3. Concerning the quality of the instruction:								
A. The instruction was excellent	2	3	4	11	4	1	1	26
B. I would consider the instruction average			1		1			2
C. The instruction was of poor quality								

latter type of conference leader presented much more practical and helpful material than the former. Therefore, The Management Institute adopted a procedure by which every conference leader is rated by the coordinator as well as by the group. The form in Figure 2 is used for the coordinator's evaluation.

This procedure in which the coordinator of the program also evaluates each conference leader was used in the 1959 ASTD Summer Institute. It was found that a coordinator's rating was usually close to the group's rating, but in some instances it varied considerably. A combination of the two ratings was used by the Management Institute staff in evaluating the effectiveness of each conference leader. In selecting and orienting future conference leaders for ASTD Institutes, both of the evaluations will be taken into consideration.

It is suggested that the training director in each company consider this approach. A trained observer such as the Training Director or another qualified person would fill out an evaluation form independent of the group's reactions. A comparison of the two would give the best indication of the effectiveness of the program.

Measuring Reactions To Outside Training Programs

The forms and suggestions that have been described above will apply best to an internal training program. Since many companies send their management people to outside training programs at universities, American Management Associations, National Industrial Conference Board, etc., it is suggested that the reaction of each person attending such a program be measured. Lowell Reed, Training Director of the Oscar Mayer & Company of Madison, Wisconsin, uses the form in Figure 3 for evaluating the reaction to the University of Wisconsin Management Institute program.

In this situation, Oscar Mayer & Company is not interested in the reaction to specific leaders. They are interested in reaction to the overall program to determine whether or not to send other foremen and supervisors. In other words, this particular questionnaire was designed to fit the need of the Oscar Mayer & Company. In addition to the tabulated responses described above, an opportunity was given each person to write in additional comments.

Another company uses the form in Figure 4 to evaluate the reaction of their managers who attend an outside program.

The first step in the evaluation process is to measure the reactions to training programs. It is important to determine how people feel about the programs they attend. Decisions by top management are frequently made on the basis of one or two comments they receive from people who have attended. A supervisory training program may be cancelled because one superintendent told the plant manager that "this program is for the birds."

Also, people must like a training program to obtain maximum benefit from it. According to Spencer, for example, "for maximum learning you must have interest and enthusiasm." In a talk given by Cloyd Steinmetz, past president of ASTD, "It is not enough to say, 'boys, here is the information, take it!' We must make it interesting and motivate them to want to take it."

To evaluate effectively, training directors should begin by doing a good job of measuring reactions and feelings of people who participate. It is important to do so in an organized fashion using written comment sheets which have been designed to obtain the desired reactions. It is also strongly suggested that the form be so designed that the comments can be tabulated and quantified. In the experience of the staff of The Management Institute, it is also desirable to have the coordinator, training director, or another trained observer make his or her own appraisal of the session in order to supplement the reactions of enrollees. The combination of these two evaluations is more meaningful than either one by itself.

Companies who send their people to attend outside institutes and conferences should make an effort to evaluate the reactions to these programs. Several suggested forms have been described.

When a training director has effectively measured the *reactions* of conferees and finds them to be very favorable, he/she can feel very proud. However, the trainer should also feel humble because the evaluation measurement has only begun.

Even though he/she has done a masterful job of measuring the reaction of the group, there is still no assurance that any learning has taken place. Neither is there any indication that the behavior of the participants will change because of the training program. And still farther away is any indication of the results that can be attributed to the training program. These three steps in the evaluation process — learning, behavior, and

Figure 4

SUPERVISORY INSTITUTE PROGRAM EVALUATION FORM

IN GENERAL

1. How worthwhile was the Institute(s) for you?
   - ☐ Very worthwhile
   - ☐ Fairly worthwhile
   - ☐ Not very worthwhile
   - ☐ A waste of time
2. Did the Institute have:
   - ☐ Too much theory and not enough of the practical
   - ☐ Too much of the practical and not enough theory
   - ☐ About the right combination of theory and practice

HOW THE INSTITUTE WAS CONDUCTED

3. On the whole, the course was conducted
   - ☐ Very well
   - ☐ Fairly well
   - ☐ Poorly
   - ☐ Very poorly
4. Lecture and discussion
   - ☐ Too much lecture
   - ☐ Too much discussion
   - ☐ About the right amount of each
5. Discussion leaders
   - ☐ Too many from the University
   - ☐ Too many from business and industry
   - ☐ O.K.
6. Visual aids
   - ☐ Not enough movies, charts, etc.
   - ☐ Too much use of demonstrations, blackboards, movies, charts, etc.
   - ☐ O.K.

APPLICATION OF THE COURSE

7. Did the Institute apply to your particular operations?
   ☐ Yes ☐ Partly ☐ No

FOLLOW-UP

8. Would you like to attend another institute?
   ☐ Yes ☐ No
   Comment
9. Should these Institutes run for ☐5 days, ☐4 days, ☐3 days.
10. Please list 3 of your main problems:
    1. ____________________
    ____
    2. ____________________
    ____
    3. ____________________
    ____
11. Comments or suggestions

results will be discussed in detail in the next three articles of this series on Evaluation.

PART 2 — LEARNING

We have emphasized in the first article that the reaction of the conferees is important in evaluating the training program. From an analysis of reactions, a training director can determine how well the program was accepted. He/she can also obtain comments and suggestions which will be helpful in improving future programs. It is important to obtain favorable reaction because:

1. Decisions on future training activities are frequently based on the reactions of one or more key persons.
2. The more favorable the reaction to the program, the more likely the conferees are to pay attention and learn the principles, facts, and techniques that are discussed.

However, it is important to recognize that favorable reaction to a program *does not assure* learning. All of us have attended meetings in which the conference leader or speaker used enthusiasm, showmanship, visual aids and illustrations to make his presentation well accepted by the group. A careful analysis of the subject content would reveal that the speaker said practically nothing of value — but did it very well. At our Management Institute at the University of Wisconsin, for example, this has been true on a number of cases. (Less and less, I hasten to add.)

Therefore, it is important to determine objectively the amount of learning that takes place. This article is aimed at suggesting ways and means for measuring this learning.

Learning Defined

There are several definitions for learning. For the purpose of this article, learning is defined in a rather limited way as follows: What principles, facts and techniques were understood and absorbed by the conferees? In other words, we are not concerned with the on-the-job use of these principles, facts, and techniques. This application will be discussed in a third article dealing with "Behavior."

Guideposts for Evaluating in Terms of Learning

Several guideposts should be used in establishing a procedure for measuring the amount of learning that takes place.

1. The learning of *each conferee* should be measured so that quantitative results can be determined.
2. A before-and-after approach should be used so that any learning can be related to the program.
3. As far as practical, the learning should be measured on an *objective* basis.
4. Where practical, a control group (not receiving the training) should be used to compare with the experimental group which receives the training.
5. Where practical, the evaluation results should be analyzed statistically so that learning can be proven in terms of correlation or level of confidence.

These guideposts indicate that evaluation in terms of learning is much more difficult than evaluation in terms of reaction as described in the first article. A knowledge of statistics, for example, is necessary. In many cases, the training department will have to call on the assistance of a statistician to plan the evaluation procedures, analyze the data, and interpret the results.

Suggested Methods

Classroom Performance: It is relatively easy to measure the learning that takes place in training programs that are teaching skills. The following programs would fall under this category: Job Instruction Training; Work Simplification; Interviewing Skills; Induction Techniques; Reading Improvement; Effective Speaking; and Effective Writing. Classroom activities such as demonstrations, individual performance of the skill being taught, and discussions following a role playing situation can be used as evaluation techniques. The training director can organize these in such a way that he/she will obtain a fairly objective evaluation of the learning that is taking place.

For example, in a course that is teaching Job Instruction Training to foremen, each foreman will demonstrate in front of the class the skills of JIT. From their performance, the training director can tell whether or not they have learned the principles of JIT and can use them, at least in a classroom situation. In a Work Simplification program, conferees can be required to fill out a "flow-process chart" and the training director can determine whether or not they know how to do it. In a Reading Improvement program, the reading speed and comprehension of each participant can be readily determined by their classroom performance. In an Effective Speaking program, each conferee is normally required to give a number of talks and an alert training director can measure to some extent the amount of learning that is taking place by observing a person's successive performances.

So in these kinds of situations, an evaluation of the learning can be built into the program. If it is organized and implemented properly, the training director can obtain a fairly objective measure of the amount of learning that has taken place. Directors can set up before - and - after situations in which each conferee demonstrates whether or not they know the principles and techniques being taught.

In every program, therefore, where skills of some kind are being taught, the training director should plan systematic classroom evaluation to measure the learning.

Paper-and-Pencil Tests: Where principles and facts are taught rather than techniques, it is more difficult to evaluate learning. The most common technique is the paper-and-pencil test. In some cases, standardized tests can be purchased to measure learning. In other cases, training directors must construct their own.

To measure the learning in human relations programs, two standardized tests are quite widely used in business and industry. The first is *How Supervise?* by File and Remmers. This is published by The Psychological Corporation of New York and has been used by a number of companies on a before-and-after basis to measure the learning that takes place. A newer test is the *Supervisory Inventory on Human Relations* by Kirkpatrick and Planty.[1] Sample test items from the latter are listed below (answered by circling "A" for agree or "DA" for disagree).

There are also standardized tests available in such areas as Creativity and Economics. In following the guideposts that were suggested in the beginning of this article, this kind of a standardized test should be used in the following manner:

1. The tests should be given to all conferees prior to the program.
2. If possible, it should also be given to a control group which is comparable to the experimental group.

PLEASE ANSWER ALL STATEMENTS EVEN IF YOU ARE NOT SURE

Copyright 1978 by D.L. Kirkpatrick

1.	Anyone is able to do almost any job if he/she tries hard enough.	A DA
2.	Intelligence consists of what we've learned since we were born.	A DA
3.	If a supervisor knows all about the work to be done, he/she is therefore qualified to teach others how to do it.	A DA
4.	We are born with certain attitudes and there is little we can do to change them.	A DA
5.	A supervisor should not praise members of the department when they do a good job because they will ask for a raise.	A DA
6.	A well trained working force is a result of maintaining a large training department.	A DA
7.	A supervisor would lose respect if he/she asked employees to help solve problems that concern them.	A DA
8.	In making a decision, a good supervisor is concerned with his/her employes' feelings about the decision.	A DA
9.	The supervisor is closer to his/her employes than to management.	A DA
10.	The best way to train a new employe is to have him/her watch a good employe at the job.	A DA
11.	Before deciding on the solution to a problem, a list of possible solutions should be made and compared.	A DA
12.	A supervisor should be willing to listen to almost anything the employes want to tell him/her.	A DA

3. These pretests should be analyzed in terms of two approaches: In the first place, the total score of each person should be tabulated. Secondly, the responses to each item of the inventory should be tabulated in terms of right and wrong answers. This second tabulation not only enables a training person to evaluate the program but also gives some tips on the knowledge and understanding of the group prior to the program. This means that in the classroom, the trainer can stress those items most frequently misunderstood.

4. After the program is over, the same test or its equivalent should be given to the conferees and also to the control group. A comparison of pretest and posttest scores and responses to individual items can then be made. A statistical analysis of this data will reveal the effectiveness of the program in terms of learning.

One important word of caution must be made. Unless the test or inventory accurately covers the material presented, it will not be a valid measure of the effectiveness of the learning. Frequently a standardized test will cover only part of the material presented in the course. Therefore, only that part of the course covered in the inventory is being evaluated. Likewise, if certain items on the inventory are not being covered, no change in these items can be expected.

There are also many examples where training directors and others responsible for programs have developed their own paper-and-pencil tests to measure learning in their programs. For example, the American Telephone and Telegraph Company has incorporated into their "Personal Factors in Management" program, a short test measuring the sensitivity and empathy. First, each individual is asked to rank in order of importance, 10 items dealing with human relations. They are then assigned to a group which is asked to work 15 minutes at the task of arriving at a group ranking of the 10 statements. Following this 15-minute "heated discussion," each individual is asked to complete a short inventory which includes the following questions:

1. A. Were you satisfied with the performance of the group?
   Yes ☐ No ☐
   B. How many will say that they were satisfied with the performance of the group?__________
2. A. Do you feel that the discussion was dominated by two or three members?
   Yes ☐ No ☐
   B. How many will say that they thought the discussion was dominated by two or three members?__________
3. A. Did you have any feelings about the items being ranked that, for some reason, you felt it wise not to express during the discussion?
   Yes ☐ No ☐
   B. How many will say that they had such feelings?__________
4. A. Did you talk as often as you wished to during the discussion?
   Yes ☐ No ☐
   B. How many will say that they talked as often as they wished to?__________

The successive class sessions then attempt to teach each conferee to be more sensitive to the feelings and ideas of other people. Later in the course, another "empathy" test is given to see whether there is an increase in sensitivity.

At an ASTD Summer Institute in Mȧdison, Wisconsin, Dr. Earl Planty of the University of Illinois introduced a test on decision making.[2] Several items from that test follow:

1. If my boss handed to me a well done piece of work and asked me to make changes on it, I would
   - ☐ prove to him/her that the job is better without changes.
   - ☐ do what he/she says and point out where he/she is wrong.
   - ☐ complete the changes without comment.

☐ request a transfer from his/her department.

2. If I were office manager and one of the best clerks kept complaining about working conditions, I would
   - ☐ try to determine the basis for the complaints.
   - ☐ transfer the person to some other section.
   - ☐ point out to the person that complaints are bad for morale.
   - ☐ ask the person to write out the complaints for your superior.
3. If my supervisor criticized my work, I would
   - ☐ compare my record with coworkers for him/her.
   - ☐ explain the reason for poor performance to him/her.
   - ☐ ask the supervisor why he/she selected me for criticism.
   - ☐ ask him/her for suggestions on how to improve.
4. If I were setting up a new procedure in an office, I would
   - ☐ do it on my own without enlisting anyone's aid.
   - ☐ ask my superiors for suggestions.
   - ☐ ask the people who work under me for suggestions.
   - ☐ discuss it with my friends who are outside of the company.

This test or one like it can be given before and after a program on decision making to determine whether or not the participants learned the principles and procedures taught in the course.

In Morris A. Savitt's article called "Is Management Training Worthwhile?"[3] he described a program that he evaluated. He devised a questionnaire which was given at the beginning of the program "to determine how much knowledge of management principles and practices the conferees had at the beginning." At the end of the 10-week program, the same questionnaire was administered to test the progress made during the course. This is an example of a questionnaire tailored to a specific program.

Daniel M. Goodacre III, formerly of the B.F. Goodrich Company, has done a great deal of work in this area. He has developed and used achievement tests which are given before and after training programs to determine the amount of learning.

And so we see that the paper-and-pencil test can be used effectively in measuring the learning that takes place in a training program. It should be emphasized again that the approach to this kind of evaluation should be systematic and statistically oriented. A comparison of before and after scores and responses can then be made to prove how much learning has taken place.

Nile Soik of the Allen-Bradley Company described an additional evaluation procedure in an article in the *Training and Development Journal.* Not only did he use the *Supervisory Inventory on Human Relations* before and after the program, but he also administered it six months later. He was measuring the forgetting that took place in the period following the program.

Summary

It is easy to see, then, that it is much more difficult to measure *learning* than it is to measure *reaction* to a program. A great deal of work is required in planning the evaluation procedure, in analyzing the data that is obtained, and interpreting the results. Wherever possible, it is suggested that training directors devise their own methods and techniques. As has been pointed out in this article, it is relatively easy to plan classroom demonstrations and presentations to measure learning where the program is aimed at the teaching of skills.

Where principles and facts are the objectives of the training program, it is advisable to use a paper-and-pencil test. Where suitable standardized tests can be found, it is easier to use them. In many programs, however, it is not possible to find a standardized test and the training person must use their skill and ingenuity in devising their own measuring instrument.

If training directors can prove that a program has been effective in terms of learning as well as in terms of reaction, they have objective data to use in selling future programs and in increasing their status and position in the company.

PART 3 — BEHAVIOR

In the two previous articles in this series, we talked about techniques for evaluating training programs in terms of (1) REACTION and (2) LEARNING. It was emphasized that in our evaluations, we can borrow techniques but we cannot borrow results.

A personal experience may be the best way of starting this third article dealing with changes in behavior. When I joined The Management Institute of the University of Wisconsin in 1949, one of my first assignments was to sit through a one-week course on "Human Relations for Foremen and Supervisors." During the week I was particularly impressed by a foreman named Herman from a Milwaukee company. Whenever a conference leader asked a question requiring a good understanding of human relations principles and techniques, Herman was the first one who raised his hand. He had all the answers in terms of good human relations approaches. I was very much impressed and I said to myself "If I were in industry, I would like to work for a man like Herman."

It so happened that I had a first cousin who was working for that company. And oddly enough, Herman was his boss. At my first opportunity, I talked with my cousin, Jim, and asked him about Herman. Jim told me that Herman may know all the principles and techniques of human relations, but he certainly does not practice them on the job. He performed as the typical "bull-of-the-woods" who had little consideration for the feelings and ideas of his subordinates.

At this time I began to realize there may be a big difference between knowing principles and techniques and using them on the job.

Robert Katz, Professor at Dartmouth, wrote an article in the July-August 1956 issue of the *Harvard Business Review.* The article was called "Human Relations Skills Can Be Sharpened." And he said: "If a person is going to change their job behavior, five basic requirements must exist":

1. They must want to improve.
2. They must recognize their own weaknesses.
3. They must work in a permissive climate.
4. They must have some help from someone who is interested and skilled.
5. They must have an opportunity to try out the new ideas.

It seems that Katz has put his finger on the problems that exist in a transition between learning and changes in behavior on the job.

Evaluation of training programs in terms of on the job behavior is more difficult than the reaction and learning evaluations described in the two previous articles. A more scientific approach is needed and many factors must be considered. During the last few years a number of attempts have been made and more and more effort is being put in this direction.

Several guideposts are to be followed in evaluating training programs in terms of behavioral changes:

1. A *Systematic* appraisal should be made of on-the-job performance on a *before-and-after* basis.
2. The appraisal of performance should be made by one or more of the following groups (the more the better):
   A. The person receiving the training
   B. Their superior or superiors
   C. Their subordinates
   D. Their peers or other people thoroughly familiar with their performance.
3. A statistical analysis should be made to compare before and after performance and relate changes to the training program.
4. The post-training appraisal should be made three months or more after the training so that the trainers have an opportunity to put into practice what they have learned. Subsequent appraisals may add to the validity of the study.
5. A control group (not receiving the training) should be used.

Some of the best evaluation studies are briefly described below.

The Fleishman-Harris Studies[4]

To evaluate a training program that had been conducted at the Central School of The International Harvester Company, Fleishman developed a study design and a battery of research instruments for measuring the effectiveness of the training. Seven paper-and-pencil questionnaires were used and the trainees, their superiors, and their subordinates were all surveyed.

To supplement the data that Fleishman had discovered. Harris conducted a follow-up study in the same organization. He used a before-and-after measure of job performance and worked with experimental and control groups. He obtained information from the trainees themselves as well as from their subordinates.

Survey Research Center Studies[5]

The Survey Research Center of the University of Michigan has contributed much to evaluation of training programs in terms of on-the-job behavior. To measure the effectiveness of a human relations program conducted by Dr. Norman Maier at the Detroit Edison Co., and to measure the results of an experimental program called "feedback," a scientific approach to evaluation was used. A basic design was to use a before-and-after measure of on-the-job performance with experimental as well as control groups. The supervisors receiving the training as well as their subordinates were surveyed in order to compare the results of the research. The instrument used for measuring these changes was an attitude and opinion survey designed and developed by the Survey Research Center.

The Lindholm Study[6]

This study was carried out in the home office of a small insurance company during the period of October, 1950 to May, 1951. A questionnaire developed as part of the research program of the Industrial Relations Center of the University of Minnesota was used. It was given on a before-and-after basis to the subordinates of those who took the training. No control group was used. A statistical analysis of the before-and-after results of the attitude survey determined the effectiveness of the program in terms of on-the-job behavior.

The Blocker Study[7]

A different approach was used in the study conducted in an insurance company having approximately 600 employees. Fifteen supervisors who took a course on "Democratic Leadership" were analyzed during the three-month period following the course. Eight of the supervisors were classified as democratic and seven were classified as authoritarian based on their behavior prior to the program.

During the three-month period immediately following the program, the changes in behavior of the supervisors were analyzed through a study of their interview records. They used standard printed forms which made provision for recording the reason for the interview, attitude of the employee, comments of the supervisor, and action taken, if any. Each supervisor was required to make a complete record of each interview. They did not know that these records were to be used for an evaluation study. There were a total of 376 interviews with 186 employees.

The interview records were classified as authoritarian or democratic. The changes in interview approach and techniques were studied during the three month period following the course to determine if on-the-job behavior of the supervisors changed.

The Tarnopol Approach[8]

In his article called "Evaluate Your Training Program," Tarnopol suggests the approach to use as well as a specific example of an evaluation experiment. He believes in the employee attitude survey given on a before-and-after basis using control as well as experimental groups. He stresses that "in our experience, five employees is a good minimum for measuring the behavior of their supervisor." He also stresses that "although canned questionnaires are available, it is advisable to use measuring instruments that are specifically suited to the requirements of both your company and your training program."

In his employee attitude approach, Tarnopol has suggested inserting some neutral questions which do not relate to the training being given. This is an added factor in interpreting the results of the research.

The Moon-Hariton Study[9]

Their study was made in an Engineering Section of a department of the General Electric Company in 1956. The staff of the General Electric Company was assisted by a representative of the Psychological Corporation.

In the spring of 1958, two years after the adoption of a new appraisal and training program, a decision was made to attempt to evaluate its effectiveness. It was felt that the opinion of the subordinates about changes in the managers' attitudes and behavior would provide a better measure than what the managers themselves thought about the benefits of the program. Thus a questionnaire was designed to obtain the subordinates' views about changes in their managers. Nevertheless, it was felt that the opinions of the manager would add to the picture.

Accordingly, they were also surveyed.

The questionnaire asked the respondents to compare present conditions with what they were two years ago. In other words, instead of measuring the attitudes before and after the program, the subordinates and the managers were asked to indicate what changes had taken place during the last two years.

Buchanan-Brunstetter Study[10]

At the Republic Aviation Corporation, an attempt was made to measure the results of a training program. The questionnaire was used and an experimental and a control group were measured. The experimental group had received the training program during the past year while the control group was going to receive it during the following year. The subordinates of the supervisors in each one of these groups were asked to complete a questionnaire which related to the on-the-job behavior of their supervisor. After answering the questionnaire in which they described the job behavior of their supervisor, they were asked to go over the questionnaire again and to place a check opposite any items: "(1) which you think are *more* effectively done now than a year ago; (2) which you think are *less* effectively done now than a year ago."

In this experiment as well as in the Moon-Hariton approach, the subordinates were asked to indicate what changes in behavior had taken place during the last year. This was done because a before measure of their behavior had not been made.

The Stroud Study[11]

A new training program called "Personal Factors in Management" was evaluated at the Bell Telephone Company of Pennsylvania by Peggy Stroud. Several different approaches were used to compare the results and obtain a more valid indication of on-the-job behavioral changes that resulted from the program. The first step was the formulation of a questionnaire to be filled out by four separate groups: (1) conferees (2) controllees (supervisors not taking the course) (3) superiors of the conferees (4) superiors of the controllees.

The first part of the questionnaire was the "Consideration Scale" taken from the leader behavior description questionnaire originated in the Ohio State leadership studies. The second part of the questionnaire was called the Critical Incident section in which the conferee and control groups were asked to describe four types of incidents that had occurred on the job. The third and final section of the questionnaire applied to the conferees only. They were asked to rate the extent to which they felt the training course had helped them achieve each of its five stated objectives.

It was decided to conduct an extensive evaluation of the training program after the program had begun. Therefore it was not possible to make a before and after comparison. In this study, an attempt was made to get the questionnaire respondents to compare on-the-job behavior before the program with that following the program. According to Miss Stroud, it would have been better to measure behavior prior to the program and then compare it to behavior measured after the program.

This study, called "Evaluating A Human Relations Training Program," is one of the best attempts this writer has discovered. The various evaluation results are compared and fairly concrete interpretations made.

The Sorenson Study[12]

The most comprehensive research that has been done to evaluate the effectiveness of a training program in terms of on-the-job behavior was made at the Crotonville Advanced Management Course of the General Electric Company. It was called the "Observed Changes Enquiry."

The purpose of the "enquiry" was to answer these questions:

1. Have manager graduates of General Electric's Advanced Management Courses of 1956 been observed to have changed in their manner of managing?

2. What inferences may be made from similarities and differences of changes observed in graduates and non-graduates?

First of all, the managers (graduates and non-graduates alike) were asked to indicate changes they had observed in their own manner of managing during the previous 12 months. Secondly, subordinates were asked to describe changes they had observed in the managers during the past 12 months. Thirdly, their peers (looking sideways) were asked to describe changes in behavior. And finally, the superiors of the control and experimental groups were asked to describe the same changes in behavior. This gave Sorensen an excellent opportunity to compare the observed changes of all four groups.

In this extensive research, Sorensen used experimental as well as control groups. He also used four different approaches to measure observed changes. These include the person, their subordinates, their peers, and their superiors. In this research, he did not use a before-and-after measure but rather asked each of the participants to indicate what changes, if any, had taken place during the past year.

Summary and Conclusions

The purpose of this article has been to describe briefly some of the best experiments that have been used to measure effectiveness of training programs in terms of on-the-job behavior. Only the methods and instruments used in these studies have been mentioned. The results, although interesting, cannot be borrowed by other training directors but the techniques can.

For those interested in evaluating in terms of behavioral changes, it is strongly suggested that these studies be carefully analyzed. The references following this series of articles indicate where the detailed articles can be found.

Once more I would like to emphasize that the future of training directors and their programs depends to a large extent on their effectiveness. To determine effectiveness, attempts must be made to measure in scientific and statistical terms. This article, dealing with changes in behavior resulting from training programs, indicates a very complicated procedure. But it is worthwhile and necessary if training programs are going to increase in effectiveness and their benefits made clear to top management.

It is obvious that very few training directors have the background, skill and time to engage in extensive evaluations. It is therefore frequently necessary to call on statisticians, research people, and consultants for advice and help.

PART 4 — RESULTS

The objectives of most training programs can be stated in terms of results desired. These results could be classified as: reduction of costs; reduction of turnover and absenteeism; reduction of grievances; increase in quality and quantity of production; or improved morale which, it is hoped, will lead to some of the previously stated results. From an evaluation standpoint, it would be best to evaluate training programs directly in terms of results desired. There are, however, so many complicating factors that it is extremely difficult if not impossible to evaluate certain kinds of programs in terms of results. Therefore, it is recommended that training directors begin to evaluate in terms of the three criteria described in the preceding articles.

First of all, determine the *reactions of* the trainees. Secondly, attempt to measure what *learning* takes place. And thirdly, try to measure the changes in on-the-job behavior. As has been stressed in the previous articles, these criteria are listed in increasing order of difficulty.

As I survey literature on evaluation, I find more and more articles being written on this subject. Nearly every issue of the ASTD *Journal* contains one or more articles. It is interesting to note that few of them deal with evaluation in terms of results. And this is because it is usually a difficult evaluation to make.

Certain kinds of training programs, though, are relatively easy to evaluate in terms of results. For example, in teaching clerical personnel to do a more effective typing job, you can measure the number of words per minute on a before and after basis.

If you are trying to reduce grievances in your plant, you can measure the number of grievances before and after the training program. If you are attempting to reduce accidents, a before and after measure can be made. One word of caution, however. E.C. Keachie stated it as follows in an issue of the ASTD *Journal:* "Difficulties in the evaluation of training are evident at the outset in the problem technically called 'the separation of variables;' that is, how much of the improvement is due to training as compared to other factors?" This is the problem that makes it very difficult to measure results that can be attributed directly to a specific training program.

Below are described several evaluations that have been made in terms of results. They do not offer specific formulas for other training directors to follow, but they do suggest procedures and approaches which can be effectively used.

Safety Programs

Many attempts have been made to evaluate the effectiveness of safety training programs in terms of lost-time accidents. One study was conducted by Philip E. Beekman, Plant Administrator of Salaried Personnel for the Colgate-Palmolive Company, Jersey City plant. This study was briefly described in the Number 3, 1958 *Supervisory Management Newsletter* of The American Management Associations.

A comparison was made of plant safety records for the nine-month period before the training program with a comparable period after the program. The frequency rate for lost-time accidents was measured along with the number of reported accidents. The frequency rate dropped from 4.5 per cent to 2.9 per cent and the number of reported accidents dropped from 41 to 32. This improvement was credited directly to the training effort because no physical changes were made which affected the accident rate.

Table 1
RESULTS OF TRAINING

	Total Number of Incidents	
Category	*Experimental Group "A"*	*Control Group "B"*
Negligent Accidents	5	8
Misdeliveries	21	33
Mishandling Valuable Mail	3	7
Late Reporting	35	32
Absence Without Reporting	3	6
Abuse of Sick Leave	8	12
Errors in Relay Operations	13	22
Adverse Probationary Reports	4	9
Discourtesy	4	5

At a 1958 Conference of The Management Institute, University of Wisconsin, Dr. G. Roy Fugal of the General Electric Company described a before-and-after evaluation of one of their safety programs. The purpose of the training was to reduce the number of accidents and to increase the regularity with which all accidents, major and minor, were reported. The training program consisted of the usual presentations, discussions, and movies which were very dramatic in describing accidents and their implications. The comprehensive evaluation indicated that the training program did not have desirable results. Therefore, a new approach to training was adopted which was more oriented to the job relationship between the foreman and each worker. An evaluation of this kind of safety training program did indicate the desired results.

Postal Carrier Training

In the September-October 1957 issue of the ASTD *Journal*, John C. Massey described a program in which he evaluated in terms of results. Experimental group "A" received 35 hours of orientation training under the post office training and development program. A comparable group called control group "B" did not receive any training. Results of this study are shown in Table 1.

The design of this evaluation study includes an experimental as well as a control group. The importance of using these was emphasized in Part 3 of this series. It should also be used in evaluating results wherever possible to overcome the difficulty described by Dr. Keachie.

An Insurance Company Study

In a recent letter, S.W. Schallert of the Farmers Mutual Insurance Company of Madison, Wisconsin reported to me on an evaluation he had made. A number of their claims adjustors were enrolled in the Vale Technical Institute of Blairsville, Pennsylvania. The purpose of the three-week course was to improve the ability of adjustors to estimate and appraise automobile physical damage.

The specific technique used by Schallert was to have the adjustors keep track of their savings for approximately six months after returning from Vale. These savings were the difference between the estimate of damage by garages and the estimate of damage by the claims adjustors who had been trained at Vale. Where the final cost of the adjustment was the same as the estimate made by the Farmers Mutual adjustor, this was considered the savings.

In other words, the purpose of the training was to prepare the adjustors to make estimates which they could justify and sell. Actual dollars and cents figures could then be used to determine whether or not the cost of sending these adjustors to Vale was justified.

A Cost Reduction Institute

Several years ago, two graduate students at the University of Wisconsin attempted to measure the results of a "Cost Reduction Institute" conducted by The Management Institute of the University. Two techniques were used. The first was to conduct depth interviews with some of the supervisors who had attended the course and with their immediate supervisors. The other technique was to mail questionnaires to the remaining enrollees and to their supervisors. Following is a brief summary of that study:

A. DEPTH INTERVIEWS

Interviews With Trainees

1. Have you been able to reduce costs in the few weeks that you have been back on the job?

Replies:
13 men—yes
3 men—no
2 men—noncommital or evasive
1 man—failed to answer

2. How? What were the estimated savings?

Different types of replies indicated that the 13 people who said they had made cost reductions had done so in different areas. But their ideas stemmed directly from the program, according to these trainees.

Interview of Superiors

Eight of the cost reduction actions described by the trainees were confirmed by the immediate superior and these superiors estimated total savings to be from $15,000 to $21,000 per year. The specific ideas that were used were described by superiors as well as by the trainees.

B. MAILED QUESTIONNAIRES

Questionnaires were mailed to those trainees who were not contacted personally. The results on the questionnaire were not nearly as specific and useful as the ones obtained by personal interview. The study concluded that it is probably better to use the personal interview rather than a questionnaire to measure this kind of program.

In the December 1955 issue of *The Harvard Business Review*, Willard V. Merrihue of General Electric Company and Raymond A. Katzell of Richardson, Bellows, Henry and Company described a very complex approach. According to them, "measuring performance is essential if we are to know whether the planning, the organizing, and all the other functions which preceded logically and time wise are, in fact, being discharged as well as they could or should be."

The ERI is designed to measure the extent to which groups of employees accept and perform in accordance with the objectives and policies of the company. The following indicators constituted the ERI: (1) periods of absence; (2) separations; (3) initial visits to the dispensary for occupational reasons; (4) suggestions submitted through the suggestion system; (5) actions incurring disciplinary suspension; (6) grievances submitted through the formal grievances procedures; (7) work stoppages; and (8) participation in the insurance plan.

At the time this article was written by Merrihue and Katzell, the ERI was in its preliminary stages. Also, it did not deal directly with evaluating training programs although it indicated it might be used as a measurement yardstick. The article in its entirety should be read by those persons who are interested in the complex area of measuring training programs in terms of results. Several practical ideas might be obtained which will be helpful in establishing specific evaluation criteria and procedures.

Measuring Organizational Performance

Another sophisticated and penetrating article related to evaluation was written by Rensis Likert. It appears in the March-April 1958 issue of *The Harvard Business Review*. It shows how changes in productivity can be measured on a before and after basis. Two different types of groups were used; the first was a group of supervisors trained in using a democratic kind of leadership in which decision making involved the participative technique. The supervisors in the other group were trained to make their own decisions and not ask subordinates for suggestions.

In addition to measuring the results in terms of productivity, such factors as loyalty, attitudes, interest, and work involvement were also measured. Where both training programs resulted in positive changes in productivity, the "participative" approach resulted in better feelings, attitudes, and other human relations factors.

The article described another excellent study from the University of Michigan. Dr. Likert concluded by saying that "industry needs more adequate measures of organizational performance than it is now getting."

Summary

And so we see that the evaluation of training programs in terms of "results" is progressing at a very slow rate. Where the objectives of training programs are as specific as the reduction of accidents, the reduction of grievances, and the reduction of costs, we find a number of attempts have been made. In a few of them, the researchers have attempted to segregate factors other than training which might have had an effect. In most cases, the measure on a before and after basis has been directly attributed to the training even though other factors might have been influential.

Studies like those of Merrihue-Katzell and Likert attempt to penetrate the difficulties encountered in measuring such programs as human relations, decision making, and the like. In the years to come, we will see more efforts along this direction and eventually we may be able to measure human relations training, for example, in terms of dollars and cents. At the present time, however, our research techniques are not adequate.

Conclusion

One purpose of these four articles has been to stimulate training directors to take a penetrating look at evaluation. It has been emphasized that their future and the future of their training programs depends to a large extent on their ability to evaluate and to use evaluation results.

The second objective has been to suggest procedures, methods, and techniques for evaluating training programs. A training director should begin by measuring in terms of results as described in Part 1 of this series. A second step should be to evaluate in terms of learning as described in Part 2. Part 3 suggested ways and means of evaluating in terms of on-the-job behavior which should also be attempted. And finally Part 4 has analyzed some of the problems and approaches to measuring training programs in terms of its final objective-results.

It is hoped that the training directors who have read and studied these four articles are now clearly oriented on the problems and approaches to evaluating training. As future articles on evaluation appear, we training people should carefully analyze them to see if we can borrow the techniques and procedures the writers describe.

It is also hoped that as training directors evaluate their training programs, they will describe the procedures they have used and use this magazine and others to inform others of what they have done. Progress in the evaluation of training will result if all of us will freely exchange information on objectives, methods, and criteria.

REFERENCES

1. Complimentary copy available from Dr. Donald Kirkpatrick, 4380 Continental Dr., Brookfield, WI 53005.
2. Published by Martin M. Bruce, 71 Hansen Lane, Rochelle, N.Y.
3. *Personnel*, September - October, 1957, American Management Associations.
4. Fleishman, E. A., Harris, E. F., Buntt, H.E., "Leadership and Supervision in Industry," Personnel Research Board, Ohio State University, Columbus, Ohio, 1955, page 110.
5. Mann, Dr. Floyd, "Human Relations in the Industrial Setting." Survey Research Center, University of Michigan, Ann Arbor, Mich.
6. Lindholm, T.R., "Supervisory Training and Employee Attitudes," *Journal of the ASTD*, Nov.-Dec., 1953.
7. Blocker, C.E., "Evaluation of a Human Relations Training Course," *Journal of the ASTD*, May-June, 1955.
8. Tarnopol, Lester, "Evaluate Your Training Program," *Journal of the ASTD*, Mar.-Apr., 1957.
9. Moon, C.G., Hariton, Theodore, "Evaluating an Appraisal and Feedback Training Program," *Personnel*, Nov.-Dec., 1958.
10. Buchanan, P.C. Brunstetter, P.H., "A Research Approach to Management Improvement," *Journal of the ASTD*, Jan. and Feb., 1959.
11. Stroud, P.V., "Evaluating a Human Relations Training Program," *Personnel*, Nov.-Dec. 1959.
12. Sorensen, Olav, "The Observed Changes Enquiry," Manager Development Consulting Service, General Electric Company, Crotonville, New York, May 15, 1958.

Donald L. Kirkpatrick is professor of Management Development at the University of Wisconsin-Extension, Milwaukee. A past national president of ASTD (1975), he has authored numerous articles and books.