"THE DIRECT AND INDIRECT COSTS ATTRIBUTABLE TO TRAINING HAVE RISEN DRASTICALLY, WHILE ITS CONTRIBUTION TO THE COMPANY'S BOTTOM-LINE IS BEING CHALLENGED."

EVALUATION IN FOCUS

BY JAY ALDEN

Evaluation of training has become a hot issue. Participants at training conferences are crowding into sessions dealing with evaluation. Few training periodicals are published without at least one article on evaluation, and some have devoted an entire issue to the subject. Companies are actually creating departments whose sole responsibility is training evaluation. The reasons for this movement are not difficult to understand. The marketplace has become extremely competitive and staff departments must prove their worth in order to survive. Yet the direct and indirect costs attributable to training have risen drastically, while its contribution to the company's bottom-line is being challenged.

Even with increased evaluation activity, results have been disappointing. Many evaluation efforts have come up empty, producing findings that are inconclusive, disbelieved or simply ignored. If a program evaluation effort is to be considered successful, some substantive decision about the program must be based on the evaluation data. The training program should be cut, lengthened, changed, resequenced, rescheduled, eliminated, or even retained as is *because* of the evaluation findings. Unfortunately, this is not usually the case. Training programs do change, of course — in fact, they change all the time — but the changes are rarely the direct result of evaluation findings.

What can account for the failure of evaluation projects to influence the form and substance of the training programs they are appraising? The explanation seems to lie in the evaluator's lack of focus, rather than in their lack of motivation or ability. During a typical project, evaluators will measure anything that moves, ask whatever questions are possible to ask, and report all results accurate to three decimal points. Is it any wonder that evaluation reports are ignored, even if they are read? Successful evaluation projects, on the other hand, focus directly on the management decisions being considered for the training programs.

Four factors need to be determined before a meaningful evaluation project can be designed.

1. Will management even consider making a decision about whether or not to change a program or how it should be changed?
2. What research questions will provide the data necessary to make the management decisions?
3. What level of data is practical to collect and important enough for management to use in the decision-making process?
4. What criteria will management use to make the decision?

Potential for Change

Consider the probable consequences of an evaluation report that shows major weaknesses in a program that management is either unable or unwilling to change. For the evaluator, it is a source of extreme frustration: for management, it is a source of political friction. Every evaluator should post a sign that reads, *"If they won't change it, don't evaluate it."* Each evaluation effort should be preceded by a commitment from management that the program can, in fact, change.

In certain cases, this commitment is easy to obtain. One such case is when an existing program is under indirect or informal attack — when reports show that *"sales are down,"* supervisors indicate *"they're not making or fixing them right,"* or memos or calls report that *"it's a lousy training program."* There is an especially high potential for change in a newly developed training program designed for large numbers of students or one that is deemed critical to the organization's bottom-line. In these cases, the cost consequences of an ineffective program are obvious to management and they are generally prepared to effect changes if the appropriate evaluation evidence is obtained. In such instances, there is an opportunity for the evaluation effort to influence the program.

In other cases, there is little or no opportunity for change. Obviously, if the current form of a program is nearing the end of its life cycle, or if the number of students projected for the program is so small that no additional resources will be invested in it, evaluation is pointless. Less obvious, but equally obstructive to successful evaluation is the situation in which informal indicators have produced a strong management commitment to an existing program (e.g., *"No one is complaining"; "Th e VP liked it;" "I like it"*). The commitment may, indeed, be so strong that any evaluation data to the contrary will be rejected outright. Before an evaluation is scheduled, management should be asked the question: *"What will you do if the evaluation findings show that many of the students going through the program do NOT achieve the objectives you have for them?"* If the answer suggests that there is little chance that the program will change, the evaluation effort is almost certain to fail.

Kinds of Questions

Assume that management will, in fact, consider evaluation results in making a decision about the future of a program. Next, agreement must be reached on the particular questions that will provide the data necessary to make those decisions. For example, one decision that management might be considering is simply, *"Should the program be changed?"* But on what basis will management make this decision? Effectiveness? Efficiency? Relevance? The fatal flaw of most evaluators is that they believe they must ask questions concerning all factors. Since their resources tend to be limited, they often end up with watered-down data that answers none of the questions to management's satisfaction.

Obviously, management will hope that their programs are effective, efficient and relevant. Yet decisions about whether or not to change a program may hinge on only one or two such factors, the others being of little concern or believed to be under control. *"We know it's a good program; we're just not sure if it's too long." "All we want to know is if they learn what we teach." "We know we're teaching well, but are we teaching the right things?"* The evaluator should attempt to focus the research questions on the critical factors — the ones that will decide the issues.

Concern with Effectiveness

Does the program enable students to:

☐ meet the end-of-course objectives?

☐ perform satisfactorily on the job?

Concern with Efficiency

Is there apparent waste in the program of:

☐ student time?

☐ instructor time?

☐ materials and equipment?

☐ use of facilities?

Concern with Relevance

Are the behaviors learned in the program:

☐ actually performed on the job?

☐ critical to job success?

☐ consistent with on-job practices and policies?

The decision, *"Should the program change?"* usually calls for a *"yes"* or *"no"* answer. Therefore, the research questions designed to help make this decision tend to key on the end result of the training program: Are the students successful after training? How much time does it take to train them? Are the tasks they were taught in school actually performed on the job? These types of questions do not necessarily identify specific program components in need of improvement or provide recommendations for how the improvements should be made. However, management's concern with the program might go beyond the issue of whether or not it should change. They might ask for assistance in deciding *"How should the program change?"*

A decision of this type requires an entirely different set of research questions. The required questions are diagnostic in nature and focus on the enabling factors that contribute to the program's effectiveness, efficiency, and relevance. For example:

Should the course entrance requirements be addressed?

☐ Are students who fail to meet prerequisites permitted to enter the course?

☐ Does level of student entrance ability predict school success?

Should particular training materials be revised?

☐ Can failure in end-of-course or job performance be traced to specific learning modules?

☐ Do students achieve the learning objectives of each course module?

☐ Is the content of each module technically accurate and relevant to job needs?

☐ Do excessive unplanned instructor interventions or discussions occur during use of particular materials?

☐ Are expensive or time-consuming materials worthwhile in terms of their effect on end-of-course performance?

☐ Is the language level appropriate to the target population?

Should the course administrative practices be changed?

☐ Are the instructors skilled in teaching?

☐ Do instructors deviate from planned administrative activities?

☐ Are nonteaching responsibilities of instructors excessive?

☐ Can administrative efficiency be improved, without adversely effecting end-of-course performance, by:

- increased student / instructor ratio?
- incentives for pressure for increased pace?
- small vs. large group instruction?
- reduction of student practice activity?
- pretest and bypass of modules?

Should the instructional design be altered?

☐ Does each module lead to a measurable goal?

☐ Does the sequence of modules facilitate learning?

☐ Are sufficient presentation, practice, and feedback provided?

☐ Are the instructional media appropriate to the learning tasks?

☐ Are the learning tasks and goals challenging and important to the target population?

☐ Has all content critical to job success been included?

Should the posttraining environment be modified?

☐ Are there opportunities for course graduates to immediately perform on the job, the tasks they were taught in school?

☐ Do course graduates receive on-the-job staff support for performing the tasks taught in school?

☐ Are the job resource materials accurate, complete and consistent with those used in school?

☐ Do management practices reinforce school-taught behaviors?

The tendency of many evaluators is to attempt to answer all such questions. However, the shotgun approach usually runs contrary to available time and resources. Evaluators must limit the diagnostic questions they ask to those they can practically answer. But which ones? Rather than derive the questions from their own pet learning theories, they should focus on management's options and concerns. Which enabling factors is management in a position to change? For which are there currently informal indicators of deficiency? Which, if changed, offer the greatest opportunities for improvements to the program?

Level of Data

Almost all the questions that can be asked during an evaluation effort can be answered with different levels of data. For example, the research question: *Is the content of the module relevant to job needs?* could be answered in any of the following ways:

☐ Ask the students to rate the relevance of the content (either during training or after they return to the job). High ratings would indicate that the content is relevant.

☐ Ask experienced practitioners or subject-matter experts to rate the relevance of the content. High ratings would indicate that the content is relevant.

☐ Observe either course graduates or experienced practitioners on the job. If they perform the task or apply the content, and do it the way they were taught in school, the content is relevant.

☐ Check the products or end results of people who perform the task or apply the content the way it is taught in school. If the results are satisfactory, the content is relevant.

These four levels of data can apply to almost any kind of research question:

1. *Participant Perception:* People who are attending or who have attended the course give their individual reactions to the question by ratings, rankings, or comments.

2. *Expert Opinion:* People who are considered to be subject-matter experts or master practitioners in the field give their individual reactions to the question by ratings, rankings, or comments.

3. *Measurement of Behavior:* A score is derived systematically by testing or observing the actions taken by a person or persons. The score represents the individual's proficiency or adherence to a standard in carrying out a prescribed series of actions.

4. *Measurement of End Results:* A score is derived systematically by testing or observing the product or result of actions taken by a person. The score represents the quality, speed and accuracy or adherence to a standard of the end results produced by the individual.

The level of data used to answer a research question is selected as a trade-off between practicality of measurement and importance to management. Participant Perception and Expert Opinion are generally the easiest to measure, but the objectivity of the data can be questioned. Management may choose to ignore or discount answers to the research questions derived by these means. Measurement of Behavior tends to give objective data that deals explicitly with the questions being asked. However, this measurement scheme often requires one-on-one monitoring (e.g., looking over people's shoulders with a clipboard in hand), which tends to be time-consuming and expensive.

Of the four levels of data described above, measurement of end results is frequently the optimum technique. Management is almost always concerned more with "products" than with "processes." (*"You can use any selling strategy you want, as long as you make sales;" "I don't care how you fix it as long as it gets fixed."*) Also, end results are usually easier to measure than behaviors. The performer doesn't even have to be present when the results are measured or counted. In fact, end results are often measured as part of existing accountability systems, and need only be captured from official data bases. However, there is a caution associated with data derived from measurement of end results. End results are a less direct indicator of proficiency than behavior measurements. A product may be imperfect because of faulty materials or environmental conditions, rather than the actions of the individual performer. Conversely, a product might prove to be acceptable in spite of an imperfect performance.

Whichever level of data is selected to answer a research question, management should be consulted about whether it is important enough to use in the decision-making process. In fact, the research question provided for preliminary management review should make reference to the level of data selected. *Is the content of the module relevant to job needs as measured by. . . .?*

Decision-Making Criteria

Suppose you asked questions about factors with which management is concerned, and have collected data at a level that they consider important. The evaluation findings might still be ignored. For example, assume the following kinds of data are reported:

☐ Eighty-three per cent of students attending the course achieve the end-of-course objective as measured by a performance test.

☐ The average rating of job relevance by experienced practitioners is 3.8 on a five-point scale, where 1 is "irrelevant" and 5 is "highly relevant."

☐ During a random sample of instructor observations, instructors deviated from planned administrative activities seven per cent of the time.

☐ On the average, 14 days elapse from the time students complete training until the opportunity arises for them to put the skill into practice.

The information appears interesting, but there is no assurance that it will be used in the decision-making process. Management may choose to ignore the data. What is missing is the minimum acceptable criterion that defines when change should or must take place. Continuing the example above, if the minimum acceptable standard for each objective was established at 80 per cent, the materials leading to the end-of-course objective that scored 83 per cent would be considered valid. On the other hand, if the standard was set at 90 per cent, the materials leading to the end-of-course objective would require revision.

The *minimum acceptable criterion* for a factor can be established in either of two ways:

• *Absolute Standard* — The criterion represents a "threshold" level that the measured factor must equal or exceed. Ideally, the standard is systematically derived from a bottom-line target, but it may result from a theoretical "model" performance or a calculated "normalized" figure — or even a "cross-your-fingers" type of expectation.

• *Comparative Standard* — The criterion is derived from the performance of another group called the "control" group. The group being measured as part of the evaluation is compared statistically with the equivalent measures on the control group. To meet the minimum acceptable criterion, the difference between the evaluation and control group measures must exceed chance probability. The control group may be composed of people who (1) were trained in the same course at an earlier time or at a different location, (2) were trained in a different course, (3) were not trained at all, or (4) are the same people in the evaluation group just before they received the training.

The advantage of the Absolute Standard is that the decision-making process requires minimal analysis techniques (i.e., the threshold level is either met or not met). Its disadvantage is that the particular threshold level selected may be difficult to justify (*"Why is 90 per cent the minimum acceptable level?"*). In comparison, the control group used with the Comparative Standard is rarely arbitrary or abstract, but the analysis technique used to reach the decision requires inferential statistics.

Whichever type of standard is used for the minimum acceptable criterion, it is advisable that it be established before any data is collected. That is, *"the money should be put on the table before the hand is played."* If at all possible, management should take part in defining the standard. At the very least, they should review and commit to its use in the decision-making process.

The introduction of minimum acceptable criteria permits the writing of research questions that can be answered "yes" or "no."

☐ Can at least 80 per cent of the students attending the course correctly diagnose and repair typical field faults as measured by a hands-on performance test administered at the end of school?

☐ Do students in the Selection Interviewing program improve their interviewing skills as measured by a performance checklist in a role-play situation?

☐ Do more graduates of the Basic Sales course attain 100 per cent of plan after three months in the field than Sales Reps trained on the job?

☐ Does the average rating of Job Relevance by students in the program exceed 3.5 on a five-point scale, where 1 is "irrelevant" and 5 is "highly relevant"?

In this form, the questions suggest a clear-cut evaluation design (i.e., what should be measured, how it should be measured, who should be measured, how results should be analyzed). But most important, they offer the greatest opportunity for findings to influence the program being evaluated. The questions focus on the issues that management cares about, and in a form in which a "no" answer is a compelling reason to effect change.

Jay Alden is the manager of Training Evaluation at the Xerox International Center for Training and Management Development, in Leesburg, Va.