

CHAPTER 11

EVALUATION IN INSTRUCTIONAL DESIGN: THE IMPACT OF KIRKPATRICK'S FOUR-LEVEL MODEL

Walter Dick
Florida State University

Editors' Introduction

Evaluation is an essential component of models of instructional design. In this chapter, Walter Dick focuses on how the evaluation process has changed (and remained the same!) over the past forty years, paying particular attention to Kirkpatrick's model of evaluation.

In the 1960s, a distinction was made between two types of evaluation: formative and summative. In the first half of this chapter, this distinction and the unique features of both evaluation processes are discussed. Moreover, one model of program evaluation, an approach to evaluation that was developed in the 1970s, is described.

The second half of the chapter focuses on Kirkpatrick's four-level model of evaluation, a model that was developed over forty years ago but that seems to be particularly relevant today. Dick describes how this model can be used for formative as well as summative purposes and indicates how its emphasis on the performance of workers and on organizational results fits in very well with the performance technology movement.

Knowledge and Comprehension Questions

1. *What major event brought to light the need for formative evaluation?*
2. *What is the major distinction between formative evaluation and summative evaluation?*
3. *Describe the four levels of Kirkpatrick's evaluation model.*
4. *Explain the similarity and difference between applying Kirkpatrick's model to an instructional design project and a performance technology project.*

P sychologists and training personnel with roots in the empirical sciences were primarily responsible for developing the original instructional design models. One of the fundamental components of these models is evaluation. The collection of data and information on the effectiveness of newly created instruction was critically important in the verification of the design process and the revision of the instruction. The purpose of this chapter is to describe the original role of evaluation in the design process, how that has evolved in practice in business and industry, and how Kirkpatrick's model of evaluation has influenced the evaluation process.

Origins of Evaluation in Instructional Design

The evaluation of educational innovations in the 1950s and 1960s usually consisted of research designs that involved the use of experimental and control groups. A posttest was used to determine whether the experimental group that received the instruction did significantly better than the control group, which had received no instruction. This approach was used to determine the effectiveness of new instructional innovations such as educational television and computer-assisted instruction. In these studies, the effectiveness of instruction delivered via the innovation was compared to the effectiveness of "traditional instruction," which was usually delivered by a teacher in a classroom. The major purpose of the evaluation was to determine the value or worth of the innovation that was being developed.

In the 1960s, the United States undertook a major curriculum reform. Millions of dollars were spent on new textbooks and approaches to instruction. As the new texts were published, the traditional approach to evaluation was invoked—namely, comparing student performance using the new curricula with the performance of students who used the traditional curricula. Although some of the results were ambiguous, it was clear that many of the students who had used the new curricula had learned very little. Several leaders in the field of educational psychology and evaluation, including Lee Cronbach and Michael Scriven, recognized that the problems with this instruction should have been discovered sooner. The debate that followed resulted in the reconceptualization of educational evaluation and the emergence of the terms *formative evaluation* and *summative evaluation*.



Formative and Summative Evaluation

The result of the discussions about the role of evaluation was the agreement that some form of evaluation had to be undertaken before distribution of textbooks to users. The purpose of the evaluation was not to determine the value or worth of the texts, but rather to determine how they could be improved before being released. During this evaluation phase, there is an interest in how well students are learning and how they like the instruction. But this information is used to make the instruction more effective, not to make a decision about how good it is. Thus the evaluation is formative in that it takes place during the development of the instruction. In contrast, summative evaluation takes place after the developers have done all that they can to make the instruction as effective as possible. During the summative evaluation, it is appropriate to ask questions such as “Is this new instruction as good as our old form of instruction? Is this web-based version of the course as effective as the CD-ROM version? Which one requires more time to complete? Which one is more expensive? Which one do learners prefer?”

Instructional design models, which were first published in the late 1960s and early 1970s, all had an evaluation component. Most referred to the formative/summative distinction and suggested that designers engage in some process in which learners studied drafts of instructional materials, where data were obtained on their performance on tests and their reactions to the instruction. This information and data were to be used to inform the revision process.

The evaluation processes that were described in the early instructional design models incorporated two unique features. The first is that testing should focus on the objectives that have been stated for the instruction. This is referred to as *criterion-referenced testing* or *objective-referenced testing*. The argument is made that the assessment instruments for systematically designed instruction should focus on the skills that the learners have been told will be taught in the instruction. The purpose of testing is not to sort out the learners in order to assign grades, but rather to determine the extent to which each objective in the instruction has been mastered. Instruction for objectives for which assessments indicate low performance should be reviewed and revised. Therefore, the assessments—be they multiple-choice items, essays, or products developed by the learners—should require learners to demonstrate the skills as they are described in the objectives in the instruction.

The second feature of evaluation within instructional design is the focus on learners as the source of data to be used for decision making about the instruction. Although subject-matter experts are typically members of the instructional design team, they cannot always accurately predict which instructional strategies will be effective and which will not. Formative evaluation in instructional design should include a subject-matter specialist's review and that of an editor, but the major source of input to the process is the learner. Formative evaluation focuses on the learners' ability to learn from the instruction and to enjoy it. The work of Kirkpatrick (1996), which was initially published in 1959, has extended the range of variables that should be of interest to the instructional designer.

 **Models of Program Evaluation**

New evaluation models continued to be developed throughout the 1970s. These models were to have a profound impact on how designers would use and value the evaluation process. These new models differed from the earlier model that employed experimental designs for post-innovation assessments. The new models were used on projects that included extensive development work, multiple organizations and agencies, and multiple forms of instructional delivery. These projects tended to have large budgets, many staff members, and were often housed in universities. The projects had multiple goals that would be achieved over time. Examples would be teacher corps projects that were aimed at reforming teacher education, modern math projects that attempted to redefine what and how children learned about mathematics.

These new projects often employed new models of evaluation. Perhaps the most influential model of that era was that of Stufflebeam (1971). This model was referred to as the CIPP Model. The acronym stands for Context, Input, Process, and Product. Context indicated the requirement to assess the environment in which an innovation would be used to determine the need for the innovation and the factors in the environment that will affect the success of its use. This context analysis is typically called a needs assessment. Stufflebeam's model indicated that the evaluator should be present from the beginning of the project and should assist in the conduct of the needs assessment and in the interpretation of the results of that assessment.

The second step in the CIPP evaluation model is Input. Questions are raised about the resources that will be used to develop the innovation. What people, funds, space, and equipment will be available for the project? Will these be sufficient to produce the desired results? Once again, the evaluator plays a key role in assessing the input to the project and analyzing the adequacy of the resources to meet the identified need.

The third and fourth steps in the CIPP model are Process and Product. These two steps correspond more closely to the kinds of evaluation that were being done by designers under the labels of formative and summative evaluation. Process evaluation examined both the ways in which the innovation was being developed and the initial effectiveness and revisions of the innovation. Data were collected, on a regular basis during the project, to inform the project leader of the current status of the project and how the innovation was being revised to meet the needs of the implementation context. The product evaluation resembled summative evaluation in that assessments were made of the success of the innovation in the target environment.

Several things should be noted about the CIPP model. It changed dramatically the involvement of the evaluator in the development process. No longer did the evaluator appear at the end of a project to design the assessment instruments and administer a pretest/posttest or experimental/control group design study. The evaluator was now a full-time member of the project team. Similarly, evaluation was not something that just happened at the end of the project; rather, it was a process that continued throughout the life of the project. The purpose of nearly all evaluation was evolving to one of assisting in the improvement of the products being produced by the project, rather than determining the absolute usefulness or effectiveness of those products.

While the use of models of evaluation was changing both in instructional design and large-scale projects (which might be using instructional design methods to develop the innovation), there was relatively little change in evaluation procedures being used in business and industry. As academics wrote about the new evaluation models, industry tended to continue to use learner responses to post-instruction attitude questionnaires as the major or only source of evaluation data about the effectiveness of their training. This was the case in spite of the fact that in 1959, Kirkpatrick published a much more extensive evaluation model for use in business and industry. The model is described in the next section.

Kirkpatrick's Four-Level Model of Evaluation

Kirkpatrick's model was published initially in four articles in 1959 (Kirkpatrick, 1996). His purpose for proposing the model was to stimulate training directors to increase their efforts to evaluate their training programs. What he originally referred to as steps later became the four levels of evaluation, but they are not necessarily interdependent. Each of the four levels is described below. As you read, remember that the distinction between formative and summative evaluation would not be made for another decade. As described by Kirkpatrick, the levels essentially refer to what we would now call summative evaluation. The statements that appear in quotes are all taken from Kirkpatrick's original articles as they were reprinted in *Training and Development* in January 1996 (see the complete reference at the end of the chapter).

Level 1: Reactions

The first level of evaluation is the assessment of learners' reactions or attitudes toward the learning experience. Questionnaires are the instruments that are used to get honest reactions from the learners so that information can be provided to management about the instruction. These reactions, along with those of the training director, should be used to evaluate the instruction but should not serve as the only type of evaluation. It is assumed that if learners do not like the instruction, it is unlikely that they will learn from it.

Level 2: Learning

The assessment of learning includes what "principles, facts, and techniques were understood and absorbed by trainees." Kirkpatrick recommended that objective means be used to measure learning and that a control group also be measured to compare their performance with that of the trainees that received the instruction. A pre-test/posttest design was suggested as a means of using the statistical techniques of the day to demonstrate that learning had occurred as a result of the instruction. Kirkpatrick states that the tests should cover the material that was presented to the learners in order to have a valid measure of the amount of learning that has taken place. (It was assumed in 1959 that the training would be classroom based. Very little non-instructor delivered instruction

was then available.) The value of having both reaction and learning data was that training directors would have a basis for “selling future programs and increasing their status in the company.”

Level 3: Behavior

Kirkpatrick found from his experience that even though someone could demonstrate learning in the classroom setting, there is no guarantee that the person will demonstrate those same skills in the job setting. Therefore the training director should do a follow-up evaluation several months after the training to determine whether the skills that were learned in the classroom are being used on the job. To determine whether the skills are being used, and how well, it is necessary to contact not only the learner, but also the learner’s supervisors, peers, and subordinates. To have a valid indication of use of skills, Kirkpatrick again suggests the use of pre- and post-assessments, including a control group to compare with the trainees. He recognizes that this level of evaluation is much more difficult than those done in the classroom, but the information is much more important to decision makers.

Level 4: Results

Kirkpatrick notes that the objectives of most training programs can be stated in terms of the desired results, such as reduced costs, higher quality, increased production, and lower rates of employee turnover and absenteeism. The behaviors acquired by learners should result in changes in the organization in the directions noted above. In essence, the need for the training is based on the failure to, or opportunity to, improve the performance of the organization. Kirkpatrick acknowledges the difficulty of being able to validate the relationship between the performance of certain learned skills and changes in the performance of an entire organization. There are too many other factors besides the training that may be influencing the changes that are observed. However, it was Kirkpatrick’s hope that training directors would attempt to improve level 4 evaluations and thus enhance the status of training organizations.

Implementation of Kirkpatrick’s Levels in Business and Industry

When Kirkpatrick reviewed the model over thirty years after it was originally published, he noted that it has remained essentially unchanged. The four levels of reaction, learning, behavior, and results remain the same. More methods have been proposed for each level, but none have changed substantially.

Are Kirkpatrick’s levels of evaluation being implemented in business and industry? The American Society for Training and Development publishes an annual report on training operations in the United States. In ASTD’s 1999 report (Bassi & Van Buren, 1999), what they refer to as “leading edge” companies were identified on the basis of the percentage of employees trained, training expenditures per employee, use of four innovative training practices, use of six high-performance work practices, and use of seven innova-

tive compensation practices. Each leading edge company was asked to identify the percentage of courses offered during 1997 that employed each of the four levels of Kirkpatrick's model. The results were as follows:

- Reaction (level 1): 81%
- Learning (level 2): 40%
- Behavior (level 3): 11%
- Results (level 4): 6%

These data indicate that in these exemplary organizations, over 80% of all courses offered are accompanied by questionnaires to determine learner attitudes toward the learning experience. The percentage falls off to 40% of courses that use a posttest to determine whether the participants have achieved the course goals. Fewer than 12% of the courses include a follow-up by designers or evaluators to determine whether the skills learned in the training are used on the job and whether their use is having the desired impact on the organization. These data indicate that even though Kirkpatrick's ideas have had wide appeal in the training community, there is little evidence that the total model is being applied to the evaluation process.

Implications of Kirkpatrick's Model for Instructional Design

Kirkpatrick's model can be applied to both instructional design and the broader concept of performance technology. Kirkpatrick's model has essentially been interpreted as a summative one. It is typically applied after training is completed to determine reactions, learning, and subsequently behaviors and results in order to validate the work of the training organization and to be persuasive with top management in the future. However, Kirkpatrick does note that his model can also be used to improve subsequent offerings of the training organization.

Dick and Carey (1996) have noted that Kirkpatrick's level 1 and 2 assessments are the same as the questionnaire and posttest approaches that instructional designers have used for several decades with various drafts of their instruction. These data are used as the fundamental information in the formative evaluation; that is, they indicate what problems learners had with the instruction and suggest what changes might be made to improve it.

Furthermore, it is possible to view level 3 and 4 evaluations from the formative point of view as well. It is invaluable for designers to determine whether the skills learned in training are being used in the performance context, and if not, why not? What are the implications for the improvement of the training? Interviews with supervisors, peers, and subordinates will indicate the extent of use and effectiveness of the new skills. Likewise, the impact of using the skills must be ascertained. Are they having the desired effect? Are they affecting sales, reducing costs, and so on? If not, how can the training be changed to provide the skills that will have the desired effect? It is clear that

Kirkpatrick's four levels of evaluation are just as useful to the instructional designer as they are to the training manager. Information from all four levels can be used to indicate the current effectiveness of the instruction and how it can be improved.

Implications of Kirkpatrick's Model for Performance Technology

More and more designers are becoming performance technologists. This means that they begin projects with a performance analysis to determine the gap in the organization's goals and its accomplishments. These gaps are examined to determine their causes, and solutions are identified that are responsive to these causes. Performance technologists often find that training, when it is required at all, often is only a part of a total solution to an organizational problem. Thus, a team is required to design, develop, implement, and evaluate the total solution.

Kirkpatrick's model is consistent with the performance technology approach. Certainly, designers will want to measure the attitudes and learning outcomes for participants in any training they develop as part of the solution to a performance problem. It is also necessary to determine whether the newly learned skills are being used on the job, along with the other components of the solution, such as improved technology or changes in procedures. Designers will also want to determine whether the implementation of the total solution is having the desired impact on the organization. The solution should solve the problem that led to the development of the solution. Thus there is a direct fit with the four levels of Kirkpatrick's model and the evaluation of the solution to an organization's performance problem.

Conclusion

In summary, evaluation has always been an essential component of the instructional design process. Kirkpatrick's model of evaluation expands the application of formative evaluation to the performance or job site. The model also is consistent with the major concepts of performance technology that are used to solve human performance problems within organizations. However, data indicate that the training departments in the best training organizations are not consistently conducting the full range of evaluations and thus are losing the benefit of this valuable information. It will be up to the designers of the future to rectify this situation.

References

- Bassi, L., & Van Buren, M. (1999, May). The 1999 ASTD state of the industry report. *Training and Development Magazine, Supplement*, 53 (5).
- Dick, W., & Carey, L. (1996). *The systematic design of instruction* (4th ed.). New York: Harper-Collins.

Kirkpatrick, D. (1996). Great ideas revisited. *Training and Development*, 50(1), 54–59.

Stufflebeam, D. L. (1971). *Educational evaluation and decision making*. Itasca, IL: F.E. Peacock Publishers.

Application Questions

1. Recent research indicates that most companies do level 1 (reaction) evaluations but relatively few do level 4 (results). Provide several explanations of why companies do fewer evaluations at the higher levels, and indicate the possible consequences of this decision.
2. Identify a recent instructional design or performance technology project on which you have worked. If you have not worked on any such project, interview someone who has. Describe how you did (or would) apply Kirkpatrick's four-level model. Indicate any problems you might have with the various levels of application.