SECTION III Evaluating and Managing Instructional Programs and Projects

Chapter 10

Evaluation in Instructional Design: A Comparison of Evaluation Models

R. Burke Johnson
University of South Alabama

Walter Dick
Florida State University

One of the fundamental components of instructional design models is evaluation. The purpose of this chapter is to describe several of the most influential and useful evaluation models.

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 if 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 learning with the new curricula with the learning of students who used the traditional curricula. While some of the results were ambiguous, it was clear that many of the students who used the new curricula 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 approach to instruction should have been discovered sooner. The debate that followed resulted in a bipartite reconceptualization of educational evaluation, and the coining of the terms *formative* and *summative* evaluation by Michael Scriven in 1967. Here are Scriven's (1991) definitions of formative and summative evaluation:

> *Formative* evaluation is evaluation designed, done, and intended to support the process of improvement, and normally commissioned or done by, and delivered to, someone who can make improvements. *Summative* evaluation is the rest of evaluation: in terms of intentions, it is evaluation done for, or by, any observers or decision makers (by contrast with developers) who need evaluative conclusions for any reasons besides development. (p. 20)

The result of the discussions about the role of evaluation in education in the late 1960s and early 1970s was an agreement that some form of evaluation needed to be undertaken prior to the distribution of textbooks to users. The purpose was not to determine the overall value or worth of the texts, but rather to determine how they could be improved. During this developmental or formative evaluation phase, there is an interest in how well students are learning and how they like and react to the instruction. Instructional design models, which were first published in

the 1960s and early 1970s, all had an evaluation component. Most included the formative/summative distinction and suggested that designers engage in some process in which drafts of instructional materials are studied by learners and data are obtained on learners' performance on tests and their reactions to the instruction. This information and data were to be used to inform revisions.

The evaluation processes described in early instructional design models incorporated two key features. First, testing should focus on the objectives that have been stated for the instruction. This is referred to as *criterion-referenced* (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 the learners to assign grades, but rather to determine the extent to which each objective in the instruction has been mastered. 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 is a focus on the learners as the primary source of data for making decisions about the instruction. While subject matter experts (SMEs) are typically members of the instructional design team, they cannot always accurately predict which instructional strategies will be effective. Formative evaluation in instructional design should include an SME review, and that of an editor, but the major source of input to this process is the learner. Formative evaluation focuses on learners' ability to learn from the instruction, and to enjoy it.

Defining Evaluation

Before we continue with our development of evaluation in instructional design, we provide a formal definition of *evaluation*. Because of the prominence of Scriven in evaluation, we will use his definition (Scriven, 1991):

> Evaluation is the process of determining the merit, worth, and value of things, and evaluations are the products of that process. (p. 139)

By *merit* Scriven is referring to the "intrinsic value" of the evaluation object or *evaluand*. By *worth*, Scriven is referring to the "market value" of the evaluand or its value to a stakeholder, an organization, or some other collective. By *value*, Scriven has in mind the idea that evaluation always involves the making of value judgments. Scriven contends that this valuing process operates for both formative and summative evaluation.

Scriven (1980) also provides a "logic of evaluation" that includes four steps. First, select the criteria of merit or worth. Second, set specific performance standards (i.e., the level of performance required) for your criteria. Third, collect performance data and compare the level of observed performance with the level of required performance dictated by the performance standards. Fourth, make the evaluative (i.e., value) judgment(s). In short, evaluation is about identifying criteria of merit and worth, setting standards, collecting data, and making value judgments.

Models of Program Evaluation

Many evaluation models were developed in the 1970s and 1980s.[1] These evaluation models were to have a profound impact on how designers would come to use the evaluation process. 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 and many staff members, and were often housed in universities. The projects had multiple goals that were to be achieved over time. Examples were teacher corps projects aimed at reforming teacher education and math projects that attempted to redefine what and how children learned about mathematics. These projects often employed new models of evaluation. Perhaps the most influential model of that era was the CIPP model developed by Stufflebeam (1971).

Stufflebeam's CIPP Evaluation Model

The CIPP acronym stands for context, input, process, and product. These are four distinct types of evaluation, and they all can be done in a single comprehensive evaluation or a single type can be done as a stand-alone evaluation.

Context evaluation is the assessment of the environment in which an innovation or program will be used, to determine the need and objectives for the innovation and to identify the factors in the environment that will impact the success of its use. This analysis is frequently called a *needs assessment*, and it is used in making *program planning decisions*. According to Stufflebeam's CIPP model, the evaluator should be present from the beginning of the project, and should assist in the conduct of the needs assessment.

---

[1]Additional evaluation models are being developed today, and many of the older models continue to be updated. For a partial listing of important models not presented in this chapter, see Chen (1990), Patton (2008), and Stufflebeam, Madaus, & Kellaghan (2000). If space allowed, the next two models we would include are Chen's "theory driven evaluation" and Patton's "utilization focused evaluation."

The second step or component of the CIPP model is *input evaluation*. Here, evaluation questions are raised about the resources that will be used to develop and conduct the innovation/program. What people, funds, space, and equipment will be available for the project? Will these be sufficient to produce the desired results? Is the conceptualization of the program adequate? Will the program design produce the desired outcomes? Are the program benefits expected to outweigh the costs of the prospective innovation/program? This type of evaluation is helpful in making *program-structuring decisions.* The evaluator should play a key role in input evaluation.

The third step or component of CIPP is *process evaluation.* This corresponds closely to *formative evaluation.* Process evaluation is used to examine the ways in which an innovation/program is being developed, the way it is implemented, and the initial effectiveness, and effectiveness after revisions. Data are collected to inform the project leader (and other program personnel) about the status of the project, how it is implemented, whether it meets legal and conceptual guidelines, and how the innovation is revised to meet the implementation objectives. Process evaluation is used to make *implementation decisions.*

The fourth component of CIPP is *product evaluation,* which focuses on the success of the innovation/program in producing the desired outcomes. Product evaluation includes measuring the outcome variables specified in the program objectives, identifying unintended outcomes, assessing program merit, and conducting cost analyses. Product evaluation is used when making *summative evaluation decisions* (e.g., "What is the overall merit and worth of the program? Should it be continued?").

Introduction of the CIPP model to instructional design changed the involvement of the evaluator in the development process. The evaluator became a member of the project team. Furthermore, evaluation was no longer something that just happens at the end of a project, but became a formal process continuing throughout the life of a project.[2]

Rossi's Five-Domain Evaluation Model

Starting in the late 1970s and continuing to today, Peter Rossi and his colleagues developed a useful evaluation model (Rossi, Lipsey, & Freeman, 2004). According to this model, each evaluation should be tailored to fit local needs, resources, and type of program. This includes tailoring the evaluation questions (what is the evaluation purpose? what specifically needs to be evaluated?), methods and procedures (selecting those that balance feasibility and rigor), and the nature of the evaluator-stakeholder relationship (who should be involved? what level of participation is desired? should an internal or an external/independent evaluator be used?). For Rossi, the evaluation questions constitute the core, from which the rest of the evaluation evolves. Therefore, it is essential that you and the key stakeholders construct a clear and agreed upon set of evaluation questions.

The Rossi model emphasizes five primary evaluation domains. Any or all domains can be conducted in an evaluation. First is *needs assessment,* which addresses this question: "Is there a need for this type of program in this context?" A *need* is the gap between the actual and desired state of affairs. Second is program *theory assessment,* which addresses this question: "Is the program conceptualized in a way that it should work?" It is the evaluator's job to help the client explicate the theory (how and why the program operates and produces the desired outcomes) if it is not currently documented. If a program is not based on sound social, psychological, and educational theory, it cannot be expected to work. This problem is called *theory failure.*[3] Third is *implementation assessment*, which addresses this question: "Was this program implemented properly and according to the program plan?" If a program is not properly operated and delivered, it has no chance of succeeding. This problem is called *implementation failure.*

The fourth evaluation domain is synonymous with the traditional social science model of evaluation, and the fifth domain is synonymous with the economic model of evaluation. The fourth domain, *impact assessment*, addresses this question: "Did this program have an impact on its intended targets?" This is the question of cause and effect. To establish cause and effect, you should use a strong experimental research design (if possible). The fifth domain, *efficiency assessment*, addresses this question: "Is the program cost effective?" It is possible that a particular program has an impact, but it is not cost effective. For example, the return on investment might be negative, the costs might outweigh the benefits, or the program might not be as efficient as a competitive program. The efficiency ratios used in these types of analysis are explained below in a footnote.[4]

[2]The CIPP model continues to be a popular evaluation model today. For more information about this model (including model updates), as well as some of the other models discussed here, go to the Evaluation Center website at Western Michigan http://www.wmich.edu/evalctr/checklists/checklistmenu.htm#models

[3]Chen and Rossi's "Theory-Driven Evaluation" (which dates back to approximately 1980) makes program theory the core concept of the evaluation. We highly recommend this model for additional study (most recently outlined in Chen, 2005).

Kirkpatrick's Training Evaluation Model

Kirkpatrick's model was published initially in four articles in 1959. Kirkpatrick's purpose for proposing his model was to motivate training directors to realize the importance of evaluation and to increase their efforts to evaluate their training programs. Kirkpatrick specifically developed his model for *training* evaluation. What he originally referred to as *steps* later became the *four levels* of evaluation. Evaluators might only conduct evaluations at the early steps or they might evaluate at all four levels. The early levels of evaluation are useful by themselves, and they are useful in helping one interpret evaluation results from the higher levels. For example, one reason transfer of training (level 3) might not take place is because learning of the skills (level 2) never took place; likewise, satisfaction (level 1) is often required if learning (level 2) and other results (levels 3 and 4) are to occur.

Level 1: Reaction. Kirkpatrick's first level is the assessment of learners' reactions or attitudes toward the learning experience. Anonymous questionnaires should be used to get honest reactions from learners about the training. These reactions, along with those of the training director, are used to evaluate the instruction, but should not serve as the only type of evaluation. It is generally assumed that if learners do not like the instruction, it is unlikely that they will learn from it.

Although level 1 evaluation is used to study the reactions of participants in training programs, it is important to understand that data can be collected on more than just a single *overall* reaction to the program (e.g., "How satisfied were you with the training event?"). Detailed level 1 information should also be collected about program components (such as the instructor, the topics, the presentation style, the schedule, the facility, the learning activities, and how engaged participants felt during the training event). It also is helpful to include open-ended items (i.e., where respondents respond in their own words). Two useful open-ended items are (1) "What do you believe are the three most important weaknesses of the program?" and (2) "What do you believe are the three most important strengths of the program?" It is usually best to use a mixture of open-ended items (such as the two questions just provided) and closed-ended items (such as providing a statement or item stem such as "The material covered in the program was relevant to my job" and asking respondents to use a four-point rating scale such as: very dissatisfied, dissatisfied, satisfied, very satisfied). Kirkpatrick (2006) provides several examples of actual questionnaires that you can use or modify for your own evaluations. The research design typically used for level 1 evaluation is the one-group posttest-only design (Table 10.1).

Level 2: Learning. In level 2 evaluation, the goal is to determine what the participants in the training program learned. By *learning* Kirkpatrick (2006) has in mind "the extent to which participants change attitudes, improve knowledge, and/or increase skill as a result of attending the program" (p. 20). Some training events will be focused on *knowledge*, some will focus on *skills*, some will focus on *attitudes*, and some will be focused on a combination of these three outcomes.

Level 2 evaluation should be focused on measuring what specifically was covered in the training event and on the specific learning objectives. Kirkpatrick emphasizes 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. Knowledge is typically measured with an *achievement test* (i.e., a test designed to measure the degree of knowledge learning that has taken place after a person has been exposed to a specific learning experience); skills are typically measured with a *performance test* (i.e., a testing situation where test takers demonstrate some real-life behavior such as creating a product or performing a process); and attitudes are typically measured with a *questionnaire* (i.e., a self-report data-collection instrument filled out by research participants designed to measure, in this case, the attitudes targeted for change in the training event).

The one-group pretest-posttest design is often sufficient for a level 2 evaluation. As you can see in Table 10.1, this design involves a pretest and posttest measurement of the training group participants on the outcome(s) of interest. The estimate of learning improvement is then taken to be the difference between the pretest and posttest scores. Kirkpatrick appropriately recommends that a control

[4]In business, financial results are often measured using the *return on investment* (ROI) index. ROI is calculated by subtracting total dollar costs associated with the program from total dollar benefits (this difference is called *net benefits*); then dividing the difference by total dollar costs, and multiplying the result by 100. An ROI value greater than zero indicates a positive return on investment. A *cost-benefit analysis* is commonly used with governmental programs; this relies on the *benefit-cost ratio*, which is calculated by dividing total dollar benefits by total dollar costs. A benefit-cost ratio of 1 is the break-even point, and values greater than 1 mean the benefits are greater than the costs. Because it can be difficult to translate benefits resulting from training and other interventions into dollar units (e.g., attitudes, satisfaction), *cost-effectiveness analysis* is often used rather than cost-benefit analysis. To calculate the *cost-effectiveness ratio* the evaluator translates training program costs into dollar units but leaves the measured benefits in their original (nondollar) units. A cost-effectiveness ratio tells you how much "bang for the buck" your training provides (e.g., how much improvement in job satisfaction is gained per dollar spent on training).

TABLE 10.1 Research designs commonly used in training evaluation

Design Strength	Design Depiction				Design Name
1. Very weak			X	O_2	One-group posttest-only design
2. Moderately weak		O_1	X	O_2	One-group pretest-posttest design
3. Moderately strong		O_1	X	O_2	Nonequivalent comparison-group design
		O_1		O_2	
4. Very strong	RA	O_1	X	O_2	Pretest-posttest control-group design
	RA	O_1		O_2	

*Note that X stands for the treatment (i.e., the training event), O_1 stands for pretest measurement, O_2 stands for posttest measurement, and RA stands for random assignment of participants to the groups. Design 3 has a control group but the participants are not randomly assigned to the groups; therefore the groups are, to a greater or lesser degree, "nonequivalent." Design 4 has random assignment and is the gold standard for providing evidence for cause and effect. For more information on these and other research designs, see Johnson and Christensen, 2010.

group also be used when possible in level 2 evaluation because it allows stronger inferences about causation. In training evaluations, this typically means that you will use the nonequivalent comparison-group design shown in Table 10.1 to demonstrate that learning has occurred as a result of the instruction. Learning data are not only helpful for documenting learning; they are also helpful to training directors in justifying their training function in their organizations.

Level 3: Behavior (Transfer of Training). Here the evaluator's goal is to determine whether the training program participants change their on-the-job behavior (OJB) as a result of having participated in the training program. Just because learning occurs in the classroom or another training stetting, there is no guarantee that a person will demonstrate those same skills in the real-world job setting. Thus, the training director should conduct a follow-up evaluation several months after the training to determine whether the skills learned are being used on the job.

Kilpatrick (2006) identifies five environments that affect transfer of training: (1) preventing environments (e.g., where the trainee's supervisor does not allow the trainee to use the new knowledge, attitudes, or skills), (2) discouraging environments (e.g., where the supervisor discourages use of the new knowledge, attitudes, or skills), (3) neutral environments (e.g., where the supervisor does not acknowledge that the training ever took place), (4) encouraging environments (e.g., where the supervisor encourages the trainee to use new knowledge, attitudes, and skills on the job), and (5) requiring environments (e.g., where the supervisor monitors and requires use of the new knowledge, attitudes, and skills in the work environment).

To determine whether the knowledge, skills, and attitudes are being used on the job, and how well, it is necessary to contact the learners and their supervisors, peers, and subordinates. Kirkpatrick oftentimes seems satisfied with the use of what we call a *retrospective survey design* (asking questions about the past in relation to the present) to measure transfer of training. A retrospective survey involves interviewing or having trainees and their supervisors, peers, and subordinates fill out questionnaires several weeks and months after the training event to measure their perceptions about whether the trainees are applying what they learned. To provide a more valid indication of transfer to the workplace, Kirkpatrick suggests using designs 2, 3, and 4 (shown in Table 10.1). Level 3 evaluation is usually much more difficult to conduct than lower level evaluations, but the resulting information is important to decision makers. If no transfer takes place, then one cannot expect to have level 4 outcomes, which is the original reason for conducting the training.

Level 4: Results. Here the evaluator's goal is to find out if the training leads to "final results." Level 4 outcomes include any outcomes that affect the performance of the organization. Some desired organizational, financial, and employee results include reduced costs, higher quality of work, increased production, lower rates of employee turnover, lower absenteeism, fewer wasted resources, improved quality of work life, improved human relations, improved organizational communication, increased sales, few grievances, higher worker morale, fewer accidents, increased job satisfaction, and importantly, increased profits. Level 4 outcomes are more distal than proximal outcomes (i.e., they often take time to appear).

Kirkpatrick acknowledges the difficulty of validating the relationship between training and level 4 outcomes. Because so many extraneous factors other than the training can influence level 4 outcomes, stronger research designs are needed (see designs 3 and 4 in Table 10.1). Unfortunately, implementation of these designs can be difficult and expensive. Nonetheless, it was Kirkpatrick's hope that training directors would attempt to conduct sound level 4 evaluations and thus enhance the status of training programs.

Brinkerhoff's Success Case Method

The next evaluation model presented here is more specialized than the previous models. It is focused on finding out what about a training or other organizational intervention worked. According to its founder, Robert Brinkerhoff, the success case method (SCM) "is a quick and simple process that combines analysis of extreme groups with case study and storytelling . . . to find out how well some organizational initiative (e.g., a training program, a new work method) is working" (p. 401, Brinkerhoff, 2005). The SCM uses the commonsense idea that an effective way to determine "what works" is to examine successful cases and compare them to unsuccessful cases. The SCM emphasizes the organizational embeddedness of programs and seeks to explicate personal and contextual factors that differentiate effective from ineffective program use and results. The SCM is popular in human performance technology because it works well with training and nontraining interventions (Surry & Stanfeld, 2008).

The SCM follows five steps (Brinkerhoff, 2003). First, you (i.e., the evaluator) focus and plan the success case (SC) study. You must identify and work with stakeholders to define the program to be evaluated, explicate its purpose, and discuss the nature of the SC approach to evaluation. You must work with stakeholders to determine their interests and concerns, and obtain agreement on the budget and time frame for the study. Finally, this is when the study design is constructed and agreed upon.

Second, construct a visual *impact model*. This includes explicating the major program goals and listing all impacts/outcomes that are hoped for or are expected to result from the program. The far left side of a typical depiction of an impact model lists "capabilities" (e.g., knowledge and skills that should be provided by the program); these are similar to Kirkpatrick's level two learning outcomes. The far right depicts "business goals" that are expected to result from the program; these are similar to Kirkpatrick's level four results outcomes. The middle columns of a typical impact model include behaviors and organizational and environmental conditions that must be present to achieve the desired business goals. These might include critical actions (i.e., applications of the capabilities) and/or key intermediate results (e.g., supervisory, environmental, and client outcomes). An impact model is helpful for knowing what to include in your questionnaire to be used in the next step.

Third, conduct a survey research study to identify the best (i.e., success) cases and the worst cases. Unlike most survey research, responses are *not* anonymous because the purpose is to identify individuals. Data are collected from everyone if there are fewer than 100 people in the population; otherwise, a random sample is drawn.[5] The survey instrument (i.e., the questionnaire) is usually quite short, unless you and the client decide to collect additional evaluation information.[6] Two key questions for the questionnaire are the following: (a) "To what extent have you been able to use the [insert name of program here] to achieve success on [insert overall business goal here]," (b) "Who is having a lot of success in using the [insert program name]?," and (c) "Who is having the least success in using the [insert program name]?" The survey data can be supplemented with performance records and any other information that might help you to locate success cases (e.g., word of mouth, customer satisfaction reports).

Fourth, schedule and conduct in-depth interviews (usually via the telephone for approximately forty-five minutes per interview) with multiple *success cases*. Sometimes you will also want to interview a few nonsuccess cases. The purpose of the fourth step is to gain detailed information necessary for documenting, with empirical evidence, the success case stories. During the interviews you should discuss categories of successful use and identify facilitating and inhibiting use factors. During the success case interviews, Brinkerhoff (2003) recommends that you address the following information categories:

a. What was used that worked (i.e., what information/strategies/skills, when, how, with whom, and where)?
b. What successful results/outcomes were achieved, and how did they make a difference?
c. What good did it do (i.e., value)?
d. What factors helped produce the successful results?
e. What additional suggestions does the interviewee have for improvement?

During nonsuccess case interviews, the focus is on barriers and reasons for lack of use of what was expected to be provided by the program. You should also obtain suggestions for increasing future use. During and after all interviews, it is important to obtain *evidence* and carefully document the validity of the findings.

Fifth, write-up and communicate the evaluation findings. In Brinkerhoff's words, this is where you "tell the

[5]For information on determining sample size, see Johnson and Christensen (2010) or Christensen, Johnson, and Turner (2010).

[6]Note that the *survey instrument* is not properly called "the survey." The "survey" is the research method that is implemented. Survey instruments include questionnaires (paper and pencil, web based) and interview protocols (used in-person, over the phone, or via technologies such as Skype or teleconferencing).

story." The report will include detailed data and evidence as well as rich narrative communicating how the program was successful and how it can be made even more successful in the future. Again, provide sufficient evidence so that the story is credible. Brinkerhoff (2003, pp. 169–172) recommends that you address the following six conclusions in the final report:

a. What worthwhile actions and results, if any, is the program helping to produce?
b. Are some parts of the program working better than others?
c. What environmental factors are helping support success, and what factors are getting in the way?
d. How widespread is the scope of success?
e. What is the ROI (return-on-investment) of the new program?
f. How much more additional value could be derived from the program?

Brinkerhoff emphasizes that success case evaluation results must be used if long-term and companywide success is to result. The most important strategy for ensuring employee "buy-in" and use of evaluation results and recommendations is to incorporate employee participation into all stages of the evaluation. For a model showing many of the factors that affect evaluation use, read Johnson (1998). Because of the importance of evaluation, the next and final evaluation model is constructed around the concept of evaluation use.

Patton's Utilization-Focused Evaluation (U-FE)

Evaluation process and findings are of no value unless they are *used*. If an evaluation is not likely to be used in any way, one should not conduct the evaluation. In the 1970s, Michael Patton introduced the utilization-focused valuation model (U-FE), and today it is in the fourth book edition (which is much expanded from earlier editions) (Patton, 2008). U-FE is "evaluation done for and with specific intended users for specific, intended uses" (Patton, 2008, p. 37). The cardinal rule in U-FE is that the utility of an evaluation is to be judged by the degree to which it is used. The evaluator focuses on use from the beginning until the end of the evaluation, and during that time, he or she continually facilitates use and organizational learning or any other process that helps ensure that the evaluation results will continue to be used once the evaluator leaves the organization. *Process use* occurs when clients learn the "logic" of evaluation and appreciate its use in the organization. Process use can empower organizational members.

U-FE follows several steps. Because U-FE is a participatory evaluation approach, the client and primary users will be actively involved in structuring, conducting, interpreting and using the evaluation and its results. Here are the major steps:

1. *Conduct a readiness assessment* (i.e., determine if the organization and its leaders are ready and able to commit to U-FE).
2. *Identify the "primary intended users" and develop a working relationship with them* (i.e., primary intended users are the key individuals in the organization that have a stake in the evaluation and have the ability, credibility, power, and teachability to work with a U-FE evaluator in conducting an evaluation and using the results).
3. *Conduct a situational analysis* (i.e., examine the political context, stakeholder interests, and potential barriers and supports to use).
4. *Identify the "primary intended uses"* (e.g., program improvement, making major decisions, generating knowledge, and process use or empowering stakeholders to know how to conduct evaluations once the evaluator has left).
5. *Focus the evaluation* (i.e., identify stakeholders' high-priority issues and questions).
6. *Design the evaluation* (that is feasible and will produce results that are credible, believable, valid, and actionable).
7. *Collect, analyze, and interpret the evaluation data* (and remember to use multiple methods and sources of evidence).
8. *Continually facilitate evaluation use*. For example, interim findings might be disseminated to the organization, rather than waiting for the "final written report." U-FE does not stop with the final report; the evaluator must work with the organization until the findings are used.
9. *Conduct a metaevaluation* (i.e., an evaluation of the evaluation) to determine (a) the degree to which intended use was achieved, (b) whether additional uses occurred, and (c) whether any misuses and/or unintended consequences occurred. The evaluation is successful only if the findings are used effectively.

Utilization-focused evaluation is a full approach to evaluation (Patton, 2008), but it also is an excellent approach to complement any of the other evaluation models presented in this chapter. Again, an evaluation that is not used is of little use to an organization; therefore, it is wise to consider the principles provided in U-FE.

To become an effective *utilization-focused evaluator*, we recommend that you take courses in human performance technology, leadership and management, industrial-organizational psychology, organizational development, organizational communication, and organizational behavior. If you become a utilization-focused evaluator, it will be your job to continually facilitate use, starting from the moment you enter the organization. You will attempt to facilitate use by helping transform the state of the organization so that it is in better shape when you leave than when you entered.

Conclusion

Evaluation has a long history in instructional design, and evaluation is important because (a) it is a part of all major models of instructional design, (b) it is a required skill for human performance technologists, (c) it provides a systematic procedure for making value judgments about programs and products, and (d) it can help improve employee and organizational performance. Some instructional designers will elect to specialize in evaluation and become full-time program evaluators. To learn more about evaluation as a profession, go to the website of the American Evaluation Association (http://www.eval.org/).

Stufflebeam's CIPP model focuses on program context (for planning decisions), inputs (for program structuring decisions), process (for implementation decisions), and product (for summative decisions). Rossi's evaluation model focuses on tailoring each evaluation to local needs and focusing on one or more of the following domains: needs, theory, process/implementation, impact, and efficiency. Kirkpatrick's model focuses on four levels of outcomes, including reactions, learning (of knowledge, skills, and attitudes), transfer of learning, and business results. Brinkerhoff's success case model focuses on locating and understanding program successes so that success can become more widespread in the organization. Patton's U-FE model focuses on conducting evaluations that will be *used*.

Data indicate that many training departments still are not consistently conducting the full range of evaluations. For example, only levels 1 and 2 of Kirkpatrick's model are conducted, thus eliminating the benefits of additional valuable information. It will be up to designers of the future to rectify this situation. This chapter provides some principles and models to get you started.

Summary of Key Principles

1. Evaluation is the process of determining the merit, worth, and value of things, and evaluations are the products of that process.
2. *Formative* evaluation focuses on improving the evaluation object, and *summative* evaluation focuses on determining the overall effectiveness, usefulness, or worth of the evaluation object.
3. Rossi shows that evaluation, broadly conceived, can include needs assessment, theory assessment, implementation assessment, impact assessment, and efficiency assessment.
4. Kirkpatrick shows that training evaluations should examine participants' reactions, their learning (of knowledge, skills, and attitudes), their use of learning when they return to the workplace, and business results.
5. Brinkerhoff shows that organizational profits can be increased by learning from success cases and applying knowledge gained from studying these cases.
6. It is important that evaluation findings are *used*, rather than "filed away," and Patton has developed an evaluation model specifically focused on producing evaluation use.
7. One effective way to increase the use of evaluation findings is through employee/stakeholder participation in the evaluation process.

Application Questions

1. Recent research indicates that most companies conduct level 1 evaluations, and many conduct level 2 evaluations. However, organizations infrequently conduct evaluations at levels 3 and 4. Describe several possible reasons why companies conduct few evaluations at the higher levels, and explain how you would attempt to increase the use of level 3 and 4 evaluations.
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) evaluate the project using one or more of the evaluation models explained in this chapter.
3. Using ideas presented in this chapter, construct *your own* evaluation model.

Author Information

R. Burke Johnson is a research methodologist, and he is a Professor in the Department of Professional Studies at the University of South Alabama.

Walter Dick is an Emeritus Professor of Instructional Systems, Florida State University.

References

Brinkerhoff, R. O. (2003). *The success case method: Find out quickly what's working and what's not.* San Francisco: Berrett-Koehler.

Brinkerhoff, R. O. (2005). Success case method. In S. Mathison, *Evaluation* (pp. 401–401). Thousand Oaks, CA: Sage.

Chen, H. T. (1990). *Theory-driven evaluation.* Newbury Park, CA: Sage.

Chen, H. T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2010). *Research methods and design* (11th ed.). Boston: Allyn & Bacon.

Johnson, R. B. (1998). Toward a theoretical model of evaluation utilization. *Evaluation and Program Planning: An International Journal, 21*, 93–110.

Johnson, R. B., & Christensen, L. B. (2010) *Educational research: Quantitative, qualitative, and mixed approaches* (4th ed.). Los Angeles: Sage.

Kirkpatrick, D. L. (2006). *Evaluating training programs: The four levels.* San Francisco: Berrett-Koehler.

Patton, M. Q. (2008). *Utilization-focused evaluation: The new century text.* Thousand Oaks, CA: Sage.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. M. W. (2004). *Evaluation: A systemic approach.* Thousand Oaks, CA: Sage.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.) *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.

Scriven, M. (1980). *The logic of evaluation.* Inverness, CA: Edge Press.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. D. Phillips (Eds.), *Evaluation and education: At quarter century* (pp. 19–64). Chicago: University of Chicago Press.

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

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2000). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.). Boston: Kluwer Academic.

Surry, D. W. & Stanfield, A. K. (2008). Performance technology. In M. K. Barbour & Orey (Eds.), The Foundations of Instructional Technology. Available at http://projects.coe.uga.edu/itFoundations/

Chapter 11

An Introduction to Return on Investment

Jack J. Phillips
ROI Institute

Patricia P. Phillips
ROI Institute

"Show Me the Money." There is nothing new about that statement, especially in business. Organizations of all types value their investments. What is new is the method that organizations can use to get there. While "showing the money" may be the ultimate report of value, organization leaders recognize that value lies in the eye of the beholder; therefore, the method used to show the money must also show the value as perceived by all stakeholders.

The Value Shift

In the past, program, project, or process success was measured by activity: number of people involved, money spent, days to complete. Little consideration was given to the benefits derived from these activities. Today the value definition has shifted: value is defined by results versus activity. More frequently, value is defined as monetary benefits compared with costs.

From learning and development to performance improvement, organizations are showing value by using the comprehensive evaluation process described in this chapter. Although this methodology had its beginnings in the 1970s, with learning and development, it has expanded and is now the most comprehensive and broad-reaching approach to demonstrating the value of project investments.

The Importance of Monetary Values

Monetary resources are limited. Organizations and individuals have choices about where to invest these resources. To ensure that monetary resources are put to best use, they must be allocated to programs, processes, and projects that yield the greatest return.

For example, if a learning program is designed to improve efficiencies and it does have that outcome, the assumption might be that the program was successful. But if the program cost more than the efficiency gains are worth, has value been added to the organization? Could a less expensive process have yielded similar or even better results, possibly reaping a positive return on investment (ROI)? Questions like these are, or should be, asked routinely for major programs. No longer will activity suffice as a measure of results. A new generation of decision makers is defining value in a new way.

The "Show-Me" Generation

Figure 11.1 illustrates the requirements of the new show-me generation. "Show-Me" implies that stakeholders want to see actual data (numbers and measures) to account for program or project value. Often a connection between learning and development and value is assumed, but that assumption soon must give way to the need to show an