45

Models and Methods for Evaluation

Ron Owston
York University, Toronto, Canada

CONTENTS

ABSTRACT

This chapter situates the evaluation of technology-based programs in the context of the field of general educational program evaluation. It begins with an overview of the main evaluation approaches developed for general educational programs, including Tyler's early conception of assessing attainment of program objectives, decision-making approaches, naturalistic evaluation, and Kirkpatrick's four levels for evaluating program effectiveness. Following this is an overview of commonly used technology-specific program evaluation criteria and frameworks. Strategies distilled from these two fields are then suggested for evaluating technology-based learning programs. These strategies emphasize clarifying the goal or purpose of the evaluation and determining the information needs of the intended audiences of the evaluation at the beginning of the project. This, in turn, suggests the most appropriate evaluation methodology to be used. The chapter concludes with a description of tools that can be used for analysis of evaluative data, followed by a brief discussion of the dissemination of evaluation results.

KEYWORDS

Effect size: A statistical measure of the difference between the mean of the control group and the mean of the experimental group in a quantitative research study.

Evaluation: The process of gathering information about the merit or worth of a program for the purpose of making decisions about its effectiveness or for program improvement.

Naturalistic evaluation: An evaluation approach that relies on qualitative methodology but gives evaluators freedom to choose the precise method used to collect, analyze, and interpret their data.

Web log file: A data file residing on a Web server that contains a record of all visitors to the site hosted by the server, where they came from, what links they clicked on, as well as other information.

INTRODUCTION

New technologies that have potential implications for learning are being developed almost daily: blogs, wikis, podcasting, response clickers, interactive pads and whiteboards, advanced educational games and simulations, and social websites, to name a few. Although individual teachers are always willing to pioneer the use of these technologies in their classrooms, system administrators often face the challenge of having to make informed decisions on whether these technologies should be adopted on a wider scale or integrated into curricula. The main criterion for their adoption frequently is how effective they are at improving learning. Because of the newness of the technologies, seldom do we have any compelling evidence of their effectiveness apart from anecdotal accounts of early adopters. This inevitably leads to a call for a formal evaluation of programs that employ the technology.

The goal of this chapter is to provide guidance to those charged with the evaluation of technology-based programs on how to approach the task. What is very apparent from an examination of the literature on technology program evaluation is the large gap between it and the literature on the general field of program evaluation. As will be seen from the discussion that follows, program evaluation has become a mature field of study that offers a variety of approaches and perspectives from which the evaluator can draw. Those writing about technology evaluation tend either to ignore the field or to give it only cursory attention on the way to developing their own approaches, so another goal of this chapter is to bridge the gap between these two fields. I take the position that technology-based program evaluation is a particular case of general program evaluation; therefore, the methods and tools in the program evaluation literature are equally applicable to technology evaluation. At the same time, the criteria that technology program evaluators offer can inform the more general evaluation approaches.

This chapter begins with a discussion of the field of general program evaluation and outlines some of the more influential evaluation approaches that have emerged. Following this is an overview of common technology program evaluation criteria and frameworks. Drawing from these two areas, I then suggest strategies that can be used to evaluate technology-based learning programs and describe several new data collection and analysis software tools that can help evaluators.

GENERAL PROGRAM EVALUATION MODELS

Evolution of Program Evaluation

Prior to the 1970s, educational program evaluators tended to concentrate on determining the extent to which a program met its stated objectives, a model first advocated by Tyler (1942) in a longitudinal study of schools in the 1930s. That model seemed sensible enough and served a generation or two of educators well, but during the 1960s and 1970s researchers began developing new evaluation models that went far beyond Tyler's original conception of evaluation.

The models that emerged were developed in response to the need to provide accountability for large U.S. government program expenditures in health, education, and welfare during this period. Scriven (1972) argued that evaluators must not be blinded by examining only the stated goals of a project as other program outcomes may be equally important. By implication, Scriven urged evaluators to cast a wide net in evaluating the results of a program by looking at both the intended and unintended outcomes. In fact, he went as far as advising evaluators to avoid the rhetoric around the program by not reading program brochures, proposals, or descriptions and to focus only on the actual outcomes. Scriven also popularized the terms *formative* and *summative evaluations* as a way of distinguishing two kinds of roles evaluators play: They can assess the merits of a program while it is still under development, or they can assess the outcomes of an already completed program. In practice, these two roles are not always as clearly demarcated as Scriven suggests; nonetheless, this distinction between the two purposes of evaluation is still widely drawn on today.

Suchman (1967) argued that evaluating the attainment of a program's goals is still essential, but more critical is to understand the intervening processes that led to those outcomes. He suggested that an evaluation should test a hypothesis such as: "Activity A will attain objective B because it is able to influence process C, which affects the occurrence of this objective" (p. 177). Following this reasoning, Weiss (1972) showed how a model could be developed and tested to explain how a chain of events in a teacher home visit program could lead to the ultimate objective of improving children's reading achievement. This early work led to the development of an approach known today as *theory-based evaluation*, *theory-driven evaluation*, or *program theory evaluation* (PTE). PTE consists of two basic elements: an explicit theory or model of how the program causes the intended or observed outcomes and an actual evaluation that is at least guided by the model (Rogers et al., 2000). The theory component is not a grand theory in the traditional social science sense, but rather it is a theory of change or plausible model of how a program is supposed to work (Bickman, 1987). The program model, often called a *logic model*, is typically developed by the evaluator in collaboration with the program developers, either before the evaluation takes place or afterwards. Evaluators then collect evidence to test the validity of the model. PTE does not suggest a methodology for testing the model, although it is often associated with qualitative methodology. Cook (2000) argues that program theory evaluators who use qualitative methods cannot establish that the observed program outcomes were caused by the program itself, as causality can only be established through experimental design. Generally speaking, the contribution of PTE is that it forces evaluators to move beyond treating the program as a black box and leads them to examining why observed changes arising from a program occurred.

Decision-Making Evaluation Approaches

During the same period, other evaluators focused on how they could help educational decision makers. Best known is Stufflebeam (1973), who viewed evaluation as a process of providing meaningful and useful information for decision alternatives. Stufflebeam proposed his *context*, *input*, *process*, and *product* (CIPP) model, which describes four kinds of evaluative activities. Context evaluation assesses the problems, needs, and opportunities present in the educational program's setting. Input evaluation assesses competing strategies and the work plans and budgets. Process evaluation monitors, documents, and assesses program activities. Product evaluation examines the impact of the program on the target audience, the quality and significance of outcomes, and the extent to which the program is sustainable and transferable. In essence, the CIPP model asks of a program: What needs to be done? How should it be done? Is it being done? Did it succeed? Stufflebeam also reconciled his model with Scriven's formative and summative evaluation by stating that formative evaluation focuses on decision making and summative evaluation on accountability.

Another popular approach that emerged was Patton's (1978) *utilization-focused evaluation*. Patton addressed the concern that evaluation findings are often ignored by decision makers. He probed evaluation program sponsors to attempt to understand why this is so and how the situation could be improved. From this study, he developed not so much an evaluation model as a general approach to evaluation that has only two fundamental requirements. First, he stated that relevant decision makers and evaluation report audiences must be clearly identified. Second, he maintained that evaluators must work actively with the decision makers to decide upon all other aspects of the evaluation, including such matters as the evaluation questions, research design, data analysis, interpretation, and dissemination. Patton admitted that the challenge of producing evaluation studies that are actually used is enormous but remained optimistic that it is possible and worth attempting.

Cronbach (1980), a student of Tyler, also focused on the decision-making process. His contribution was to emphasize the political context of decision making, saying that it is seldom a lone person who makes decisions about a program; rather, decisions are more likely to be made in a lively political setting by a policy-shaping community. Cronbach advocated that the evaluator should be a teacher, educating the client group throughout the evaluation process by helping them refine their evaluation questions and determine what technical and political actions are best for them. During this educative process, the evaluator is constantly giving feedback to the clients, and the final evaluation report is only one more vehicle for communicating with them. Unlike the other evaluation theorists mentioned above, Cronbach did not believe that the evaluator should determine the worthiness of a program nor provide recommended courses of action.

Naturalistic Evaluation Approaches

At the same time these researchers were developing approaches that focused on how evaluation results are used, others concentrated their efforts on developing methods that place few, if any, constraints on the evaluator. Known as *naturalistic* or *qualitative*, these approaches give the evaluator freedom to choose the

methods used to collect, analyze, and interpret their data. Stake's (1975) *responsive evaluation* is one such model. Stake was concerned that conventional approaches were not sufficiently receptive to the needs of the evaluation client. He advocated that evaluators must attend to actual program activities rather than intents, respond to the audience's needs for information, and present different value perspectives when reporting on the success and failure of a program. Stake believed that evaluators should use whatever data-gathering schemes seem appropriate; however, he did emphasize that they will likely rely heavily on human observers and judges. Rather than relying on methodologies of experimental psychology, as is often done in conventional evaluations, Stake saw evaluators drawing more from the traditions of anthropology and journalism in carrying out their studies.

Two other approaches are of interest in this discussion of naturalistic methods. First, is Eisner's (1979) *connoisseurship model*, which is rooted in the field of art criticism. His model relies on the evaluator's judgment to assess the quality of an educational program, just as the art critic appraises the complexity of a work of art. Two concepts are key to Eisner's model: *educational connoisseurship* and *educational criticism*. Educational connoisseurship involves the appreciation of the finer points of an educational program, a talent that derives from the evaluator's experience and background in the domain. Educational criticism relies on the evaluator's ability to verbalize the features of the program, so those who do not have the level of appreciation that the connoisseur has can fully understand the program's features.

The second approach is *ethnographic evaluation*, whose proponents believe can yield a more meaningful picture of an educational program than would be possible using traditional scientific methods (Guba, 1978). Ethnographic evaluators immerse themselves in the program they are studying by taking part in the day-to-day activities of the individuals being studied. Their data-gathering tools include field notes, key informant interviews, case histories, and surveys. Their goal is to produce a rich description of the program and to convey their appraisal of the program to the program stakeholders.

Kirkpatrick's Four Levels

Although it is well established in the human resource development community, Kirkpatrick's (2001) *four-level model* is less known in educational evaluation circles because it focuses on the evaluation of corporate training programs. I have placed it in a category by itself because it has little in common with the other models discussed, as Kirkpatrick does not emphasize negotiation with the decision makers nor does he favor a naturalistic approach. Kirkpatrick's first writing on the model dates back to over 40 years ago, but it was not until more recently that he provided a detailed elaboration of its features. Even though it focuses on training program evaluation, the model is still relevant to general educational settings; for example, Guskey (2000) adapted it for the evaluation of teacher professional development programs.

Kirkpatrick proposed four levels that the evaluator must attend to: *reaction*, *learning*, *behavior*, and *results*. *Reaction* refers to the program participants' satisfaction with the program; the typical course evaluation survey measures reaction. *Learning* is the extent to which participants change attitudes, improve their knowledge, or increase their skills as a result of attending the program; course exams, tests, or surveys measure this kind of change. The next two levels are new to most educational evaluators and are increasingly more difficult to assess. *Behavior* refers to the extent to which participants' behavior changes as a result of attending the course; to assess this level, the evaluator must determine whether participants' new knowledge, skills, or attitudes transfer to the job or another situation such as a subsequent course. The fourth evaluation level, *results*, focuses on the lasting changes to the organization that occurred as a consequence of the course, such as increased productivity, improved management, or improved quality. In a formal educational setting, the fourth evaluation level could refer to assessing how students perform on the job after graduation. Kirkpatrick has recommended the use of control group comparisons to assess a program's effectiveness at these two higher levels, if at all possible.

TECHNOLOGY EVALUATION APPROACHES

So far I have concentrated on models that are applicable to a wide range of educational programs, whether or not they might involve technology. Several frameworks have been proposed specifically to assess technology-based learning, although none has been employed much by researchers other than their developers. These frameworks tend to recommend areas in which evaluators should focus their data collection, provide criteria against which technology-based learning could be judged, or provide questions for the evaluator to ask. For example, Riel and Harasim (1994) proposed three areas on which data collection might focus for the evaluation of online discussion groups: the structure of network environment, social interaction that occurs

TABLE 45.1
CIAO! Framework

	Context	Interactions	Outcomes
Rationale	To evaluate technology, we need to know about its aims and the context of its use.	Observing students and obtaining process data help us to understand why and how some element works in addition to whether or not it works.	Being able to attribute learning outcomes to technology when it is one part of a multifaceted course is very difficult. It is important to try to assess both cognitive and affective learning outcomes (e.g., changes in perceptions and attitudes).
Data	Designers' and course teams' aims Policy documents and meeting records	Records of student interactions Student diaries Online logs	Measures of learning Changes in students' attitudes and perceptions
Methods	Interviews with technology program designers and course team members Analysis of policy documents	Observation Diaries Video/audio and computer recording	Interviews Questionnaires Tests

Source: Adapted from Scanlon, E. et al., *Educ. Technol. Soc.*, 3(4), 101–107, 2000.

during the course or project, and the effects of the experience on individuals. Bates and Poole's (2003) SECTION model calls for the comparison of two or more online instructional delivery modes on the basis of the appropriateness of the technology for the targeted students, its ease of use and reliability, costs, teaching and learning factors, interactivity fostered by the technology, organizational issues, novelty of the technology, and how quickly courses can be mounted and updated. Ravitz (1998) suggested a framework that encourages the assessment of a project's evolution through interactive discussion, continual recordkeeping, and documentation. Mandinach (2005) has given evaluators a set of key questions to ask about an e-learning program in three general areas: student learning, pedagogical and intuitional issues, and broader policy issues. Finally, Baker and Herman (2003) have proposed an approach, which they call *distributed evaluation*, to deal with large-scale, longitudinal evaluation of technology. They emphasize clarifying evaluation goals across all stakeholders, using a variety of quantitative and qualitative measures ranging from questionnaires and informal classroom tests to standardized tests, designing lengthier studies so changes can be assessed over time, collecting data at the local level and entering them into a systemwide repository, and providing feedback targeted at various audiences.

Of particular note because of its origins and comprehensiveness is the context, interactions, attitudes, and outcomes (CIAO!) framework developed by Scanlon et al. (2000). The CIAO! framework represents a culmination of some 25 years of technology evaluation experience of the authors at the Open University in the United Kingdom. As shown in Table 45.1, the columns in the framework represent three dimensions of the technology-based learning program that must be evaluated: the *context* dimension concerns how the technology fits within the course and where and how it is used; *interactions* refers to how students interact with the technology and with each other; and *outcomes* deals with how students change as a result of using the technology. The first row of the framework provides a brief rationale for the need to evaluate each of the three dimensions. The second and third rows, respectively, highlight the kinds of data that should be collected for each dimension and the methods that should be employed for each. The authors point out that, while the framework has proven to be very valuable in highlighting areas in which evaluative data should be collected, caution should be exercised in not applying the framework in an overly prescriptive manner.

Perhaps the most widely used criteria for evaluating teaching with technology in higher education are the *Seven Principles for Good Practice in Undergraduate Education*, described in a seminal article by Chickering and Gamson (1987). Almost 10 years after this article was published, Chickering and Ehrmann (1996) illustrated how the criteria, which were distilled from decades of research on the undergraduate education experience, could be adapted for information and communication technologies. Briefly, the criteria suggest that faculty should:

- Encourage contact between students and the faculty.
- Develop reciprocity and cooperation among students.
- Encourage active learning.
- Give prompt feedback.
- Emphasize time on task.
- Communicate high expectations.
- Respect diverse talents and ways of learning.

Graham and colleagues applied the criteria to the evaluation of four online courses in a professional school of a large midwestern American university (Graham et al., 2000). The evaluation team developed a list of "lessons learned" for online instruction, aimed at improving the courses and which correspond to the seven principles. Similarly, Cook et al. (2003a) applied the criteria to the evaluation of a technology-enhanced undergraduate economics course. They used the principles as the basis of codes for the qualitative analysis of open-ended student survey responses and assessed the extent to which the criteria were exemplified in the course.

Although the *Seven Principles* describe effective teaching from the faculty member's perspective, the American Psychological Association has produced an often-cited list of 14 principles that pertain to the learner and the learning process (see http://www.apa.org/ed/lcp2/lcp14.html). The learner-centered principles are intended to deal holistically with learners in the context of real-world learning situations; thus, they are best understood as an organized set of principles that influence the learner and learning with no principle viewed in isolation. The 14 principles, which are grouped into four main categories, are as follows:

- *Cognitive and metacognitive* (six principles): Nature of the learning process; goals of the learning process; construction of knowledge; strategic thinking; thinking about thinking; context of learning
- *Motivational and affective* (three principles): Motivational and emotional influences on learning; intrinsic motivation to learn; effects of motivation on effort
- *Developmental and social* (two principles): Developmental influences on learning; social influences on learning
- *Individual difference factors* (three principles): Individual differences in learning; learning and diversity; standards and assessment

Bonk and Cummings (1998) discussed how these principles are relevant for the design of online courses from a learner-centered perspective and for providing a framework for the benefits, implications, problems, and solutions of online instruction. By implication, the APA principles could serve as criteria to guide the evaluation of the effectiveness of technology-based learning environments.

IMPLICATIONS FOR THE EVALUATION OF TECHNOLOGY

What should be abundantly clear at this point is the surfeit of evaluation approaches, criteria, and models. Few experienced evaluators, however, pick one model and adhere to it for all of their work; they are more likely to draw upon different aspects of several models. Worthen and Saunders (1987, p. 151) expressed this well:

> The value of alternative approaches lies in their capacity to help us think, to present and provoke new ideas and techniques, and to serve as mental checklists of things we ought to consider, remember, or worry about. Their heuristic value is very high; their prescriptive value seems much less.

Several implications can be drawn from this discussion of models so far that will help in making decisions about the design of technology-based program evaluations. These are summarized in Figure 45.1. First, we must clarify why we are proposing an evaluation: Is it to assess a blended learning course developed by a faculty member who was given a course development grant? Is it to evaluate an elementary school laptop computer initiative? Is it being conducted because students are expressing dissatisfaction with an online course? Is it to see how an online professional learning community facilitates pedagogical change? The purpose of the evaluation will lead us to favor one approach over another; for example, in the case of the faculty member developing a course, the *Seven Principles* and/or the APA's learner-centered principles may be good criteria to judge the course. The *Seven Principles* may also be appropriate to guide the evaluation of the course where there is student dissatisfaction. On the other hand, in the case of the professional program, Kirkpatrick's model (or Guskey's extension of it) would direct us not only to examining teachers' perceptions of and learnings in the community but also to studying the impact of the program on the classroom practice. Table 45.2 provides additional guidance on selecting a model from among the most widely used ones for six common program evaluation purposes. Readers should exercise caution when interpreting the table, as there are no hard and fast rules about what model to use for a given purpose. Rarely is one model the only appropriate one to use in an evaluation; however, more often than not some models are better than others for a particular study.

We next have to give careful thought about who the intended audiences of the evaluation report are and should plan on providing those individuals with the

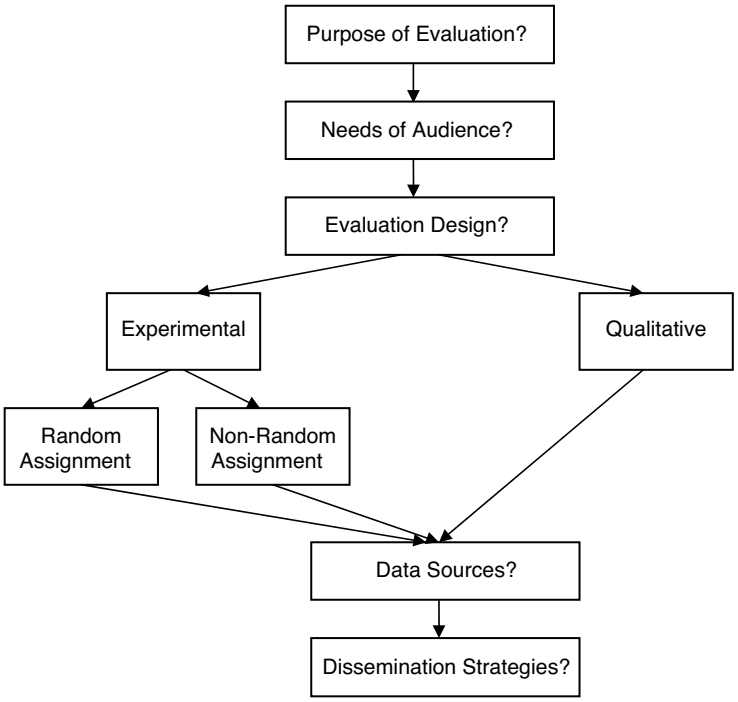

Figure 45.1 Decisions for designing an evaluation study.

TABLE 45.2
Evaluation Models Best Suited for Particular Evaluation Purposes

	Primary Purpose of Evaluation					
Evaluation Model	**Attainment of the Program's Goals and Objectives**	**Program Improvement**	**Accreditation of the Program**	**Development of Theory about Intervention**	**Meeting Information Needs of Diverse Audiences**	**Overall Impact of Program**
Goal-based (Tyler, 1942)	X	X				
Goal-free evaluation (Scriven, 1972)	X	X				X
Theory-based (Weiss, 1972)	X	X		X		X
Context, input, process, and product (CIPP) (Stufflebeam, 1973)	X	X				X
Utilization-focused (Patton, 1978)					X	
Responsive (Stake, 1975)	X	X			X	
Connoisseurship (Eisner, 1979)			X			
Ethnographic (Guba, 1978)	X	X		X		X
Multilevel (Guskey, 2000; Kirkpatrick, 2001)		X	X			X
CIAO! framework (Scanlon et al., 2000)	X	X				X
Seven principles of good practice in undergraduate education (Chickering and Ehrmann, 1996)		X				X

kinds of data needed to take appropriate action. Recall Stufflebeam's statement that the purpose of evaluation is to present options to decision makers. In a university setting, the decision makers or stakeholders might be a faculty member who is teaching an online course, a curriculum committee, a technology roundtable, a faculty

council, or senior academic administrators. The stakeholders in a school setting could be a combination of parents, teachers, a school council, and the district superintendent. The challenge to the evaluator, therefore, is to identify these audiences and then find out what their expectations are for the evaluation and the kind of information they seek about the program. Patton, Cronback, and Stake all emphasized the critical importance of this stage. The process may involve face-to-face meetings with the different stakeholders, telephone interviews, or brief surveys. Because consensus in expectations is unlikely to be found, the evaluator will have to make judgments about the relative importance of each stakeholder and whose information should be given priority.

With the expectations and information needs in hand, the study now must be planned. We saw from Scriven's perspective that all program outcomes should be examined whether or not they are stated as objectives. My experience has taught me not only to assess the accomplishment of program objectives, as this is typically what stakeholders want done, but also to seek data on unintended outcomes, whether positive or negative, as they can lead to insights one might otherwise have missed.

Design of Study

Next the evaluator must decide upon the actual design of the study. A major decision has to be made about whether to embark on an experimental design involving a comparison group or a non-experimental design. The information needs of the stakeholders should determine the path to follow (Patton, 1978; Stake, 1975). If the stakeholders seek *proof* that a technology-based program works, then an experimental design is what is likely required. The What Works Clearinghouse established by the U.S. Department of Education's Institute of Education Sciences holds experimental designs as the epitome of "scientific evidence" for determining the effectiveness of educational interventions (http://www.w-w-c.org). On the other hand, if the stakeholders seek information on how to improve a program, then non-experimental or qualitative approaches may be appropriate. Some even argue that defining a placebo and treatment does not make sense given the nature of education; hence, accumulation of evidence over time and qualitative studies are a more meaningful means of determining what works (Olson, 2004).

If a decision is made to conduct a randomized experimental study, Cook et al. (2003b) offer some helpful advice. They suggest that, rather than asking a broad question such as, "Do computers enhance learning?" (p. 18), the evaluator should formulate a more precise question that will address the incremental impact of technology within a more global experience of technology use. The authors illustrate, for example, how a study could be designed around a narrower question: "What effect does Internet research have on student learning?" (p. 19). Rather than simply comparing students who do research on the Internet with those who do not, they created a factorial design in which the presence or absence of Internet research is linked to whether teachers do or do not instruct students on best practices for Internet research. The result is four experimental conditions: best practice with Internet, best practice without Internet, typical Internet practice, and a control group whose teacher neither encourages nor discourages students from doing Internet research. The authors' recommendation echoes that offered by Carol Weiss some time ago when she made the point that the control group does not necessarily have to receive no treatment at all; it can receive a lesser version of the treatment program (Weiss, 1972). This advice is particularly relevant when speaking of technology, as it is commonly used by students today either in classrooms or outside of school, so to expect that the control group contains students who do not use technology would be unrealistic.

A problem that Cook et al. (2003b) mention only in passing is that of sample size and units of analysis—key considerations in an experimental study. In a report commissioned by the U.S. Institute of Education Sciences, Agodini et al. (2003) analyzed these issues when developing specifications for a national study on the effectiveness of technology applications on student achievement in mathematics and reading. The authors concluded that an effect size of 0.35 would be a reasonable minimum goal for such a study because previous studies of technology have detected effects of this size, and it was judged to be sufficiently large to close the achievement gaps between various segments of the student population. An effect size of 0.35 means that the effect of the treatment is 35% larger than the standard deviation of the outcome measure being considered. To achieve this effect size would require the following number of students under the given conditions of random assignment:

- *Students randomly assigned to treatments* would require 10 classrooms with 20 students in each (total of 200 students).
- *Classrooms randomly assigned to treatments* would require 30 classrooms with 20 students in each (total of 600 students) for a study of the effects of technology on reading achievement; however, 40 classrooms with 20 students (total of 800 students) would be

required for mathematics because of statistical considerations on the way mathematics scores cluster.
- *Schools randomly assigned to treatments* would require 29 schools with 20 students in each (total of 1160 students).

The first condition of random assignment of students to treatment is not likely a very feasible option in most schools, so the evaluator is left with the choice of random assignment to classrooms or to schools, both of which would require many more students. The result is that an evaluation of technology using an experimental design would likely be a fairly costly undertaking if these guidelines are followed.

Unfortunately, even random assignment to classrooms or schools may be problematic; therefore, the evaluator is left with having to compare intact classes, a design that is weak (Campbell et al., 1966). Finding teachers or students from an intact class to act as a comparison group is difficult. Even if their cooperation is obtained, so many possible competing hypotheses could explain any differences found between experimental and comparison groups (e.g., the comparison group may have an exceptional teacher or the students in the experimental group may be more motivated) that they undermine the validity of the findings.

When the goal of the study is program improvement rather than proving the program works, qualitative approaches such as those of Stake and of Guba described earlier in this chapter are particularly appropriate. Owston (2000) argued that the mixing of both qualitative and quantitative methods shows stronger potential for capturing and understanding the richness and complexity of e-learning environments than if either approach is used solely. Although some methodologists may argue against mixing research paradigms, I take a more pragmatic stance that stresses the importance and predominance of the research questions over the paradigm. This approach frees the evaluator to choose whatever methods are most appropriate to answer the questions once they are articulated. Ultimately, as Feuer et al. (2002) pointed out, "No method is good, bad, scientific, or unscientific in itself; rather, it is the appropriate application of method to a particular problem that enables judgments about scientific quality."

Data Sources and Analysis

When the basic design of the study is developed, the next decision will be to determine the evaluation data sources. Generally, the best strategy is to use as many different sources as practical, such as test scores or scores on other dependent measures, individual and focus group interviews of students and teachers, Web-based survey data, relevant program documents, and classroom observation. The use of multiple data sources is standard practice in qualitative evaluation, as the need to triangulate observations is essential (Patton, 2002). In experimental studies, other qualitative and quantitative data sources may be used to help explain and interpret observed differences on dependent measures.

Log files generated by Web servers are a relatively new source of data that can be used to triangulate findings from surveys and interviews when the technology being evaluated is Web based. These files contain a record of communication between a Web browser and a Web server in text-based form. The files vary slightly depending on the type of server, but most Web servers record the following information:

- Address of the computer requesting a file
- Date and time of the request
- Web address of the file requested
- Method used for the requested file
- Return code from the Web server that specifies if the request was successful or failed and why
- Size of the file requested

Web server log files do not reveal or record the content of a Web browser request—only the fact that a request was made. Because each Web page has a distinct address, it is possible to determine that a user viewed a particular page. Log files grow to be exceedingly large and are often discarded by system administrators; however, evaluators can analyze the files using commercial tools such as WebTrends Log Analyzer (http://www.webtrends.com) or freeware tools such as AWStats (http://awstats.sourceforge.net). Output from the tools can be in tabular or graphical format (see Figure 45.2 for sample output). The tools can be used by the evaluator to answer questions such as what time of day or week users were accessing the system, how long they were logged into the system, what pages they viewed, and what paths they followed through the website. Figure 45.2 is typical of the graphical output that may be obtained on the average number of users visiting a website per day of the week.

The author and his colleagues have used log file analysis successfully in several technology evaluation studies. In one study, Wideman et al. (1998) found that students in a focus group said they made frequent use of a simulation routine in an online course, but the log files revealed that the routine was seldom used. In another study, Cook et al. (2003a)

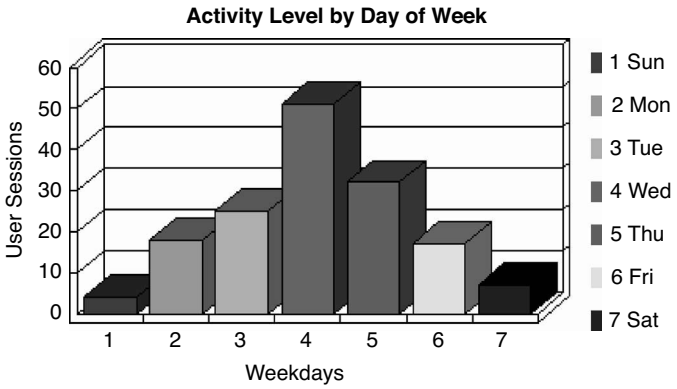

Figure 45.2 Sample output from log file analysis.

were able to correlate student access to a university course website to final course grades to obtain an indicator of how helpful the site was to students. The researchers were able to obtain these data because the website required students to log in, and a record of each log-in appeared in the log file which could be matched to the student grades. Log-file analysis has some limitations (Haigh and Megarity, 1998), but we found that it provided more and better quality data than are generated by, for example, the course management system WebCT (http://www.webct.com).

Another tool developed by the author and his colleagues to aid in the evaluation of technology-based learning is the Virtual Usability Lab (VULab) (Owston et al., 2005). VULab was originally developed for educational game research, but it is applicable to any Web-based learning research where the learner's computer is connected to the Internet. The tool allows for the automated integration of a wide range of sources of data, ranging from user activity logs, online demographic questionnaire responses, and data from automatically triggered pop-up questions (see example in Figure 45.3) to the results of queries designed to automatically appear at key points when users interact with the application. Another feature of VULab is its capability to record the screens and voice conversations of remote users and store the files on the VULab server without the need to install special software on the users' computers. The data that are collected are stored in an integrated database system, allowing for subsequent data mining and *ad hoc* querying of the data by researchers. VULab also allows for ease of use for researchers in setting up the parameters for studies and automatically monitoring users whether they are interacting with computers locally or are scattered across the Internet. Owston et al. (2005) reported on how VULab was used to record student discussions when they were filling out an online questionnaire after play-

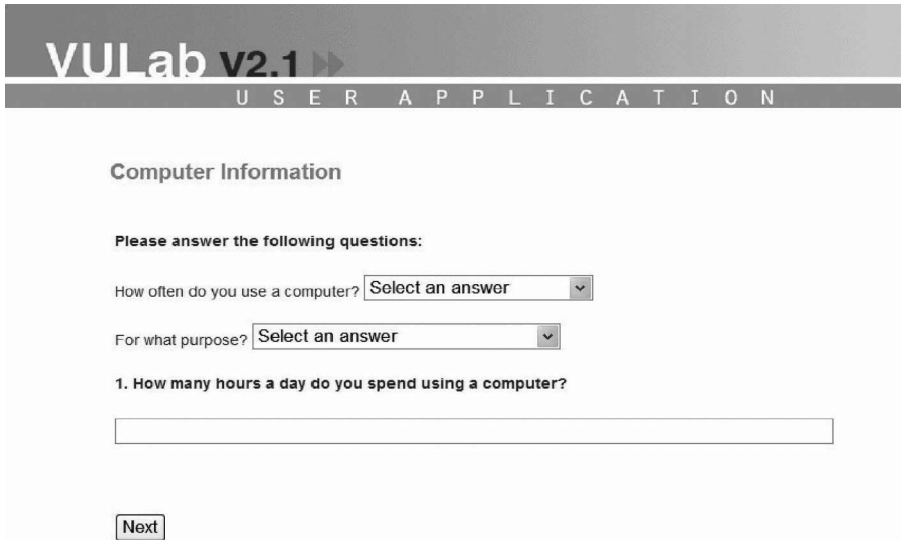

Figure 45.3 Screen shot of VULab.

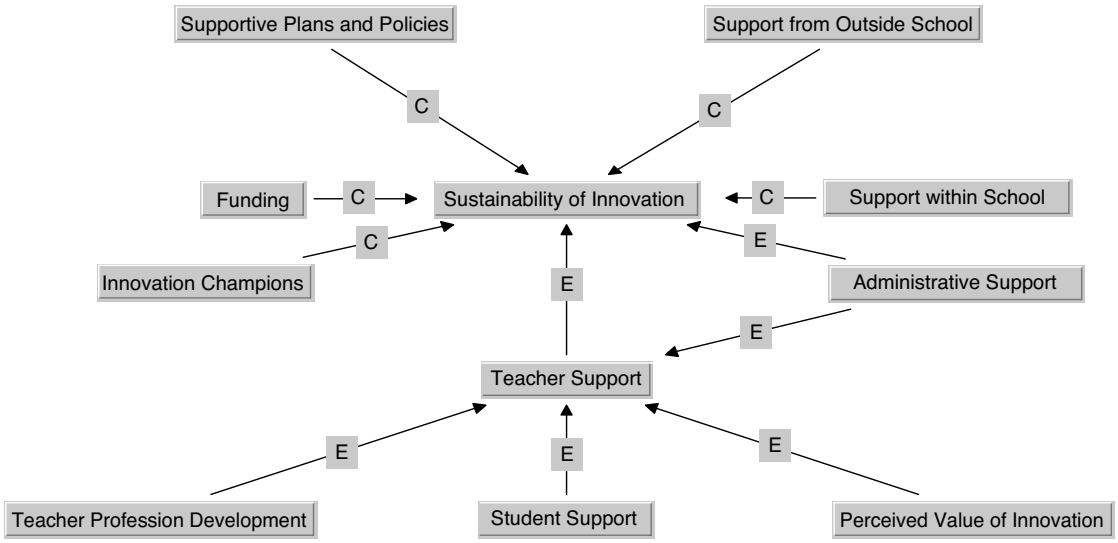

Figure 45.4 Essential (E) and contributing (C) factors to the sustainability of innovative use of technology in the classroom. (Adapted from Owston, R.D., *J. Educ. Change*, 8(1), 61–77, 2007.)

ing an online game. The students were asked on the questionnaire whether or not they enjoyed playing the game, and a rich discussion of several minutes' duration ensued among a small group of students playing the game at one computer. When it came time to enter their responses into the questionnaire form, they simply entered "yes"; thus, valuable user feedback would have been lost if it had not been for the VULab recording. The tool also proved useful for identifying role playing among the groups of students playing the game, intra-group competition and collaboration, and pinpointing technical problems within the game itself.

Frequently, evaluations involve collecting large quantities of qualitative data, such as interview transcripts, open-ended responses to questionnaires, diaries, field notes, program documents, and minutes of meetings. Managing and analyzing these files can be simplified using qualitative data analysis (QDA) software tools. Two of the most popular QDA tools are Atlas.ti (http://atlasti.com/) and NVivo (http://www.qsrinternational.com/). These tools do not perform the analysis, but they help in the coding and interpretation of the data. Both of these tools also have a feature that allows researchers to visually map relationships between codes that may lead to theory development; for example, Owston (2007) studied factors that contribute to the sustainability of innovative classroom use of technology. Using Atlas.ti, he mapped the relationships among codes and developed a model (see Figure 45.4) that helps explain why teachers are likely to sustain innovative pedagogical practices using technology. Atlas.ti allows the importing of audio and video files as well as textual files, whereas NVivo does not. In Atlas.ti, these files are coded the same way as textual files; in NVivo, the files cannot be directly imported but coding of external video and audio files can be done. If a project involves only audio or video, the best strategy may be to use Transana (http://transana.org) which is a free, open-source tool designed for the analysis of these kinds of files. A helpful feature of Transana is that while audio or video files are being played a typist can transcribe the voices directly into a separate window within the application.

An excellent website maintained by the Computer-Assisted Qualitative Data Analysis (CAQDAS) Networking Project (see http://caqdas.soc.surrey.ac.uk/) in the United Kingdom provides independent academic comparisons of popular qualitative data analysis tools and as well as other helpful resources and announcements. Those new to computerized analysis of qualitative data are well advised to visit this website for guidance in selecting the most appropriate tool to use in an evaluation.

Dissemination

A final issue that needs addressing is the dissemination of evaluation findings. The American Evaluation Association's *Guiding Principles for Evaluators* (see http://www.eval.org/Publications/GuidingPrinciples.asp) provides valuable advice to evaluators who are disseminating their results. Evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret, and critique their work. They should make clear the limitations of an evaluation and its results.

Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses significantly affecting the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.

Beyond this, the final report should contain no surprises for the stakeholders if evaluators are doing their job properly. That means that there should be an ongoing dialog between the evaluators and stakeholders, including formal and informal progress reports. This allows for the stakeholders to make adjustments to the program while it is in progress. At the same time, it is a way of gradually breaking news to the stakeholders if it looks as though serious problems are occurring with the program. Surprising stakeholders at the end of a project with bad news is one way to ensure that the evaluation report will be buried and never seen again! All the evaluation models reviewed in this chapter encourage, to varying degrees, continuous dialog between evaluators and stakeholders for these reasons. The end result should be that the evaluation report is used and its recommendations or implications are given due consideration.

CONCLUSIONS

The challenge facing evaluators of technology-based programs is to design studies that can provide the feedback needed to enhance their design or to provide evidence on their effectiveness. Evaluators need to look broadly across the field of program evaluation theory to help discern the critical elements required for a successful evaluation undertaking. These include attention to aspects such as the audience of the report and their information needs, deciding to what extent the study will be influenced by stated objectives, whether a comparative design will be used, and if quantitative, qualitative, or a combination of methods will be brought into play. The study should also be guided by the criteria and approaches developed for or applicable to the evaluation of e-learning. When these steps are taken, evaluators will be well on their way to devising studies that will be able to answer some of the pressing issues facing teaching and learning with technology.

REFERENCES

Agodini, R., Dynarski, M., Honey, M., and Levin, D. (2003). *The Effectiveness of Educational Technology: Issues and Recommendations for the National Study*, *Draft*. Washington, D.C.: U.S. Department of Education.

Baker, E. L. and Herman, J. L. (2003). Technology and evaluation. In *Evaluating Educational Technology: Effective Research Designs for Improving Learning*, edited by G. Haertel and B. Means, pp. 133–168. New York: Teachers College Press.*

Bates, A. and Poole, G. (2003). *Effective Teaching with Technology in Higher Education*. San Francisco, CA: Jossey-Bass.

Bickman, L. (1987). The functions of program theory. In *Using Program Theory in Evaluation: New Directions for Program Evaluation*, Vol. 33, edited by L. Bickman, pp. 5–18. San Francisco, CA: Jossey-Bass.*

Bonk, C. J. and Cummings, J. A. (1998). A dozen recommendations for placing the student at the centre of Web-based learning. *Educ. Media Int.*, 35(2), 82–89.

Bonk, C. J., Wisher, R. A., and Lee, J. (2003). Moderating learner-centered e-learning: problems and solutions, benefits and implications. In *Online Collaborative Learning: Theory and Practice*, edited by T. S. Roberts, pp. 54–85. Hershey, PA: Idea Group Publishing.

Campbell, D. T., Stanley, J. C., and Gage, N. L. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.*

Chickering, A. and Ehrmann, S. C. (1996). *Implementing the Seven Principles: Technology As Lever*, http://www.tltgroup.org/Seven/Home.htm.

Chickering, A. and Gamson, Z. (1987). Seven principles of good practice in undergraduate education. *AAHE Bull.*, 39, 3–7 (http://www.tltgroup.org/Seven/Home.htm).

Cook, K., Cohen, A. J., and Owston, R. D. (2003a). *If You Build It, Will They Come? Students' Use of and Attitudes towards Distributed Learning Enhancements in an Introductory Lecture Course*, Institute for Research on Learning Technologies Technical Report 2003-1. Toronto: York University (http://www.yorku.ca/irlt/reports.html).

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. *New Direct. Eval. Challenges Oppor. Program Theory Eval.*, 87, 27–34.

Cook, T. D., Means, B., Haertel, G., and Michalchik, V. (2003b). The case for using randomized experiments in research on newer educational technologies: a critique of the objections raised and alternatives. In *Evaluating Educational Technology: Effective Research Designs for Improving Learning*, edited by G. Haertel and B. Means. New York: Teachers College Press.

Cronbach, L. J. (1980). *Toward Reform of Program Evaluation*. San Francisco, CA: Jossey-Bass.*

Eisner, E. W. (1979). *The Educational Imagination: On the Design and Evaluation of School Programs*. New York: Macmillan.*

Feuer, M. J., Towne, L., and Shavelson, R. J. (2002). Scientific culture and educational research. *Educ. Res.*, 31, 4–14.

Graham, C., Cagiltay, K., Craner, J., Lim, B., and Duffy, T. M. (2000). *Teaching in a Web-Based Distance Learning Environment: An Evaluation Summary Based on Four Courses*, Center for Research on Learning and Technology Technical Report No. 13-00. Bloomington: Indiana University (http://crlt.indiana.edu/publications/crlt00-13.pdf).

Guba, E. G. (1978). *Toward a Method of Naturalistic Inquiry in Educational Evaluation*, Center for the Study of Evaluation Monograph Series No. 8. Los Angeles: University of California at Los Angeles.*

Guskey, T. R. (2000). *Evaluating Professional Development*. Thousand Oaks, CA: Corwin Press.

Haigh, S. and Megarity, J. (1998). *Measuring Web Site Usage: Log File Analysis*. Ottawa, ON: National Library of Canada (http://www.collectionscanada.ca/9/1/p1-256-e.html).

Kirkpatrick, D. L. (2001). *Evaluating Training Programs: The Four Levels*, 2 ed. San Francisco, CA: Berrett-Koehler.*

Mandinach, E. B. (2005). The development of effective evaluation methods for e-learning: a concept paper and action plan. *Teachers Coll. Rec.*, 107(8), 1814–1835.

Olson, D. R. (2004). The triumph of hope over experience in the search for 'what works': a response to Slavin. *Educ. Res.*, 33(1), 24–26.

Owston, R. D. (2000). Evaluating Web-based learning environments: strategies and insights. *CyberPsychol. Behav.*, 3(1), 79–87.*

Owston, R. D. (2007). Contextual factors that sustain innovative pedagogical practice using technology: an international study. *J. Educ. Change*, 8(1), 61–77.

Owston, R. D. and Wideman, H. H. (1999). *Internet-Based Courses at Atkinson College: An Initial Assessment*, Centre for the Study of Computers in Education Technical Report No. 99-1. Toronto: York University (http://www.yorku.ca/irlt/reports.html).

Owston, R. D., Kushniruk, A., Ho, F., Pitts, K., and Wideman, H. (2005). Improving the design of Web-based games and simulations through usability research. In *Proceedings of the ED-MEDIA 2005: World Conference on Educational, Multimedia, Hypermedia, and Telecommunications*, June 29–July 1, Montreal, Canada, pp. 1162–1167.

Patton, M. Q. (1978). *Utilization-Focused Evaluation*. Beverly Hills, CA: SAGE.*

Patton, M. Q. (2002). *Qualitative Evaluation and Research Methods*, 3rd ed. Thousand Oaks, CA: SAGE.

Ravitz, J. (1998). Evaluating learning networks: a special challenge for Web-based instruction. In *Web-Based Instruction*, edited by B. Khan, pp. 361–368. Englewood Cliffs, NJ: Educational Technology Publications.

Riel, M. and Harasim, L. (1994). Research perspectives on network learning. *Machine-Mediated Learning*, 4(2/3), 91–113.

Rogers, P. J., Hacsi, T. A., Petrosino, A., and Huebner, T. A., Eds. (2000). *Program Theory in Evaluation Challenges and Opportunities: New Directions for Evaluation, No. 87*. San Francisco, CA: Jossey-Bass.

Scanlon, E., Jones, A., Barnard, J., Thompson, J., and Calder, J. (2000). Evaluating information and communication technologies for learning. *Educ. Technol. Soc.*, 3(4), 101–107.

Scriven, M. (1972). Pros and cons about goal free evaluation. *Eval. Comm.*, 3(4), 1–7.*

Stake, R. E. (1975). *Evaluating the Arts in Education: A Responsive Approach*. Columbus, OH: Merrill.*

Suchman, E. (1967). *Evaluative Research: Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage Foundation.

Stufflebeam, D. L. (1973). An introduction to the PDK book: educational evaluation and decision-making. In *Educational Evaluation: Theory and Practice*, edited by B. L. Worthern and J. R. Sanders, pp. 128–142. Belmont, CA: Wadsworth.*

Tyler, R. W. (1942). General statement on evaluation. *J. Educ. Res.*, 35, 492–501.

Weiss, C. H. (1972). *Evaluation Research: Methods for Assessing Program Effectiveness*. Englewood Cliffs, NJ: Prentice Hall.*

Wideman, H. H., Owston, R. D., and Quann, V. (1998). *A Formative Evaluation of the VITAL Tutorial 'Introduction to Computer Science',* Centre for the Study of Computers in Education Technical Report No. 98-1. Toronto: York University (http://www.yorku.ca/irlt/reports.html).

Worthen, B. L. and Sanders, J. R. (1987). *Educational Evaluation: Alternative Approaches and Practical Guidelines*. New York: Longman.*

* Indicates a core reference.