Journal of Vocational Education and Training

Volume 15
Number 1
2007

Article Reprint

Program Evaluation

Official Publication of IVETA

Journal of Vocational Education and Training

Volume 15, Number 1

Table of Contents

Three Evaluation Models for Evaluating Program Efficiency and Effectiveness

Ernest W. Brewer, Elizabeth Peters, and John Cummins
The University of Tennessee, Knoxville

Heather Kindall
University of Arkansas, Little Rock

Abstract

This article provides a comprehensive overview of the literature related to three of the many evaluation models that can be applied to programs providing services (training, counseling, or any type of intervention) or products to students, personnel, or program participants. The three models addressed are the CIPP Evaluation Model, the Kirkpatrick Four-Step Evaluation Framework, and the Outcome-Base Evaluation Model. These models are capable of helping decision makers assess the effectiveness and efficiency of programs or projects. The following discussion is based on over 50 different sources in the literature and focuses on the advantages and disadvantages of these models.

Introduction

Hectic scheduling, tight funds, high-pressure careers, and the stress that accompanies a multitude of events, meetings, and deadlines are common factors of contemporary life for most Americans. These factors not only affect individuals as they struggle to balance family time with work while maintaining a reasonable but limited household budget, but extend into the professional world. Concerns about education, cost-effectiveness, program outcomes, accountability, and providing assistance to those in need are evident in the workforce. Royse, Thyer, Padgett, and Logan (2006) identified four main reasons programs are evaluated: required evaluation, competition for scarce funds, evaluation of new interventions, and evaluation of accountability. These areas are increasingly addressed through the implementation of program evaluations.

Defining Program Evaluation

According to Sanders and Sullins (2006), program evaluation is "the

process of systematically determining the quality of a program and how it can be improved" (p. 1). Fitzpatrick, Sanders, and Worthen (2004) concluded that program evaluations are conducted as systematic and objective processes that collect, analyze, and interpret information. More specifically, program evaluation deals with collecting and documenting information about a particular program to enable valid decision-making pertaining to a particular aspect of that program (McNamara, 2000). The ultimate purpose of program evaluation is to arrive at a definitive, intelligent, objective, and valid conclusion regarding specified objectives and questions related to a program's overall effectiveness (Fitzpatrick et al.). Education legislation, such as the No Child Left Behind Act and the 2007 Reauthorization of the Elementary and Secondary Education Act (ESEA), have increased the importance of timely and effective program evaluation.

Typically, program evaluations are used to measure progress in achieving objectives, improve program implementation, provide accountability information to stakeholders, assure funding institutions about effectiveness, increase community support for initiatives, and inform policy decisions. Posavac and Carey (2007) observed that program evaluation is "a collection of methods, skills, and sensitivities necessary to determine whether a human service is needed and likely to be used, whether the services is sufficiently intensive to meet the unmet needs identified, whether the service is offered as planned, and whether the service actually does help people in need at a reasonable cost" (p. 2). In addition, McNamara (2000) noted that improvement, in practice, implementation, and reproduction, is the goal of any high-quality program evaluation.

Uses of Program Evaluation

Program evaluation can be beneficial in a variety of contexts. It can aid in developing a concrete understanding of a program's intended outcomes and personnel requirements, or it can promote an analysis of the program's efficiency and cost-effectiveness. Furthermore, program evaluations have expanded to encompass more complex issues, no longer focusing solely on establishing cause-and-effect relationships between expectations and outcomes. Instead, they are increasingly utilized for making program decisions that relate to effectiveness, efficiency, value, and adequacy based upon a variety of systematic data collections and analyses (Rossi & Freeman, 1993). Validity is ensured via the utilization of reproducible study techniques. Program evaluations, although varied in style of implementation, must produce a basis for valid comparisons between similar programs (McNamara, 2000).

There are many different types of evaluative measures depending on the

object(s) or program(s) being assessed and the purpose of the evaluation (Fitzpatrick et al., 2004). These measures are typically distinguished by the manner in which information will be obtained and used throughout the life of the program. The appropriateness of an evaluative measure has a direct correlation to the specific nature of information that is being sought (McNamara, 2000). Ultimately, the evaluator will determine the best means of evaluation. This judgment will be based on a combination of methodologies considered, an understanding of the information needed, and knowledge from personal experiences and beliefs (Fitzpatrick et al.).

A program evaluation design is most dependent upon the information required to meet the objectives established by the group seeking the evaluation (McNamara, 2000). Consequently, a *focused evaluation* that addresses the full set of objectives of a varied group of stakeholders and other personnel will produce the most time-efficient and cost-effective evaluation (Fitzpatrick et al., 2004). In addition, the overall goal to consider when selecting an evaluation method is how to arrive at the most beneficial information to key stakeholders in the most cost-effective and realistic method (McNamara). Because evaluative measures and needs differ significantly based upon the objective, a variety of program evaluation approaches is necessary to meet the needs of different audiences and stakeholders (Stufflebeam, 1966). The three evaluation methods this article addresses that illustrate this fact are the management-oriented evaluation known as the CIPP Model, the Outcome-Based Evaluation (OBE) Model, and the four-step training evaluation Kirkpatrick's Model. Each of these models has several steps within its framework. Table 1 presents the steps associated within the framework of each model.

Table 1

Steps Associated Within the Framework of Each Evaluation Model

CIPP Evaluation Model		**Outcome-Based Evaluation (OBE)**		**Kirkpatrick Model**
Evaluation Levels	Class of Decision Making	Evaluation Levels	Methods	Framework Levels
Context	*Planning*	*Program*	*Formative*	*Reaction*
Input	*Structuring*	*Effectiveness*		*Learning*
Process	*Implementing*	*Impact*	*Summative*	*Behavior*
Product	*Recycling*	*Policy*		*Results*

Introducing the Models

CIPP Model

The CIPP Model is a management-oriented evaluation designed to provide definitive and valid information for decision makers in typical business-minded, managerial-type positions, while taking into account the different levels of decisions and decision makers. It considers who will use the evaluation results, how they will be used them, and to what aspect(s) of the system they will be applied (Fitzpatrick et al., 2004). Created as a result of appalling conditions of Chicago inner-city schools, Daniel Stufflebeam was determined to create an evaluation method that "aimed at helping people make better decisions about their educational programs and products" (EEPA Interview, 1980, p. 86). Focused on the idea that evaluative information is an essential part of good decision making, the management-oriented approach strives to provide pertinent information for those who require legitimate, unbiased evaluative information in order to arrive at a compelling judgment concerning the current state of a program (Fitzpatrick et al.). This approach further illustrates that the evaluator's effectiveness is contingent on the ability to meet the informational needs of those who are in managerial positions and responsible for implementing programs or objective results (Stufflebeam, 2003).

According to the CIPP Evaluation Model, an evaluation is defined as "a systematic investigation of the value of a program or other evaluand" (Stufflebeam, 2003, p. 9). Focusing on these values-oriented approaches, Stufflebeam further defined operational evaluation for this model as a process of "delineating, obtaining, reporting, and applying descriptive and judgmental information about some object merit, worth, probity, and significance" (p. 10). Moreover, he noted that an effective evaluation requires identifying and continually guiding a decision, providing accountability information, and advocating effective program methodologies.

Stufflebeam designed the CIPP Evaluation Model to address four different "classes" of decision making: *planning* (selective objectives), *structuring* (designing a project around specified objectives), *implementing* (operating and executing a project), and *recycling* (judgment and reaction), which all directly correlate with the evaluation methods of this model.

An operational knowledge of the four evaluative methods associated with the CIPP Evaluation Model requires an understanding of what each evaluation method or procedure entails. Although each method is designed to address specific aspects of program evaluation particularly important to decision-makers, this model is neither linear nor systematic (Fitzpatrick et al., 2004). The model can be easily modified to include only the specific information required by those in decision-making positions (Stufflebeam,

2003). This evaluation model has been enjoying increased popularity of late, surpassing Kirkpatrick's four-step training evaluation model.

Kirkpatrick's Model [Framework]

The Kirkpatrick's Evaluation Model has been the model most widely recognized and used for evaluating training programs. This model asserts that training program effectiveness can be evaluated by looking at four separate levels: *reaction* (how training participants react to the training), *learning* (the extent to which participants change attitudes, increase knowledge, and/or increase skill), *behavior* (the extent to which change in behavior occurs), and *results* (the final results of the training). Increasing amounts of information are accumulated as one proceeds from one level to another. The evaluation becomes more expensive and difficult to process with each successive level. Consequently, most evaluation occurs during the first two levels, which has been a source of concern for some evaluators (Arnold, 2003; Arthur, Gage-Little, & Munson, 2002a; Newstrom, 1978).

Kirkpatrick's Model has been used for a long time because it is simple, practical, and effective; it is also flexible and complete. Allinger and Janak (1989) observed that the power of the model is found in its simplicity and its ability to help people think about training evaluation criteria. It also provides a vocabulary and rough taxonomy for criteria.

The Kirkpatrick framework has not been without controversy. Critics see the model as going too far (Bates, 2004); not far enough (Birnbrauer, 1987; Riotto, 2004; Schumann, Anderson, Scott, & Lawton, 2001; Watkins, Leigh, Foshay, & Kaufman, 1998); having faulty assumptions (Allinger & Janak, 2004; Bates, 2004), or in need of replacement (Holton, 1996a; Nickols, 2004). Others claim that it is not a model, does not have hierarchal steps, and is not even an evaluation (Holton, 1996a, 1996b). Even Kirkpatrick says he never called it a model, but rather a framework (Kirkpatrick, 1996). Nickols (2004) proposed totally replacing the model and developing a stakeholder approach to evaluating training. Despite the opposition to Kirkpatrick's framework, evaluators continue to use the model in a wide variety of training areas such as the Navy (Rikard & Broadway, 2003), academia (Boyle & Crosby, 1997), communication technologies (Baskin, Barker, & Woods, 2005), educational simulations (Schumann et al., 2001), industrial/organizational psychology (Allinger & Janak), and e-Learning (Kruse, n.d.).

According to Boyle and Crosby (1997), measures such as student numbers, recruitment data, and retention of students in higher education are very important. However, they point out that the survival of some academic programs may be more dependent "on presenting more program-related data from a variety of sources" (p. 81). A systematic approach to the

evaluation of higher education, such as the Kirkpatrick Model, is important because it will measure such factors as student satisfaction, learning proficiency, application skills, and overall program effectiveness, which are all parameters that should be considered (Boyle & Crosby).

Outcome-based Evaluation Model

Outcome-based evaluation (OBE) encompasses the goals programs ought to achieve for the persons receiving their services and outlines the evaluation stakeholders are requesting, which are the organization-referenced outcomes that reflect the organization's effectiveness and efficiency. According to Schalock (2001), the following methods are the basis for OBE:

> A type of evaluation that uses person- and organization-referenced outcomes to determine current and desired person- and program-referenced outcomes and their use (program evaluation), the extent to which a program meets its goals and objectives (effectiveness evaluation), whether a program makes a difference compared to either no program or an alternative program (impact evaluation), or the equity, efficiency, or effectiveness of policy outcomes (policy evaluation). (p. 27)

Each of these types considers the outcomes in a different light and therefore is used to accomplish different purposes. The OBE focuses on the benefits gained by clients from using the program, and looks at whether the client's needs are met by the activities of the organization. The focus is on changes in the lives of clients, not on the program. The benefits to the client (i.e., the outcomes) may vary but often include things like increased knowledge, changes in attitudes, learned skills, and changes in condition. This type of evaluation is often used in education, health care, and social service programs, all of which want information about actual benefits for clients who use programs (i.e., demonstrating whether programs really make a difference).

As an evaluation process, the history of outcome-based evaluation has two main sources of origin. One is the passing of the Government Performance and Results Act (GPRA) in 1993. This law was been created to streamline United States government reporting of federal funds usage (OMB, 1993). The other source is the United Way's creation in 1996 of a specific evaluation process to streamline its reports by funded organizations (United Way of America, 1996). It had formulated an evaluation process that no longer focused on the service providers, but on the recipients of services. This also provided a unified reporting system, cutting down costs and time by combining many different evaluations into one.

Even though evaluation techniques have evolved over the last '40 years, there are several approaches that align closely with outcome-based evaluation. Many formative and summative approaches by evaluators such as Chambers (1994), Posavac and Carey (2007), and Scriven (1999) are very similar to the methods of OBE. *Formative methods* encompass program evaluation and effectiveness evaluation; *summative methods* align with impact evaluation and policy evaluation. The performance and efficiency measurement is similar to program evaluation and effectiveness evaluation. The theory-driven program evaluation of Chen and Rossi (1989) can be seen in the outcome-based methodological pluralism evaluation model. Finally, logic models whose focus is on accountability, managing for results and quality, and the use of evaluation results for improvement are consistent with the concept of formative feedback.

Constructional Framework of the CIPP Evaluation Model

Context Evaluation

The objective of context evaluation in the CIPP Evaluation Model is to identify initial information concerning how the program will function (Fitzpatrick et al., 2004). However, context evaluation goes beyond context definition to incorporate identification of the audience and its needs, as well as comparison of the program's intents with stakeholder requirements (Stufflebeam, 2003). This method of evaluation employs a variety of techniques such as surveys with stakeholders and project participants, system analyses, review of documents and archived data, implementation of diagnostic tests, and multiple interviews. The overall purpose of this type of evaluation is to ensure that there is a general consensus of the settings, goals, and objectives associated with the evaluation.

Input Evaluation

Input evaluation is conducted as a means of putting support systems, solution strategies, and procedural designs in place for the upcoming implementation of the program (Fitzpatrick et al., 2004). To execute this evaluation method, the evaluator must establish available resources, including personal and accessible materials (Stufflebeam, 2003). Similar or exemplary programs may be identified and examined to generate a more concise understanding of the particulars associated with the program. These particulars include issues such as costs, results, advantages, disadvantages, and other program related factors.

Process Evaluation

Process evaluation is a means of implementing and refining the program's design and working procedures. This evaluation procedure addresses information about how well the implementation of the program is going and what, if any, obstacles conflict with the program's success. Additionally, revisions are instigated to ensure that the program is not on a course toward failure. The key factor in implementing this type of evaluation and ensuring its success is interaction (Stufflebeam, 2003). Interaction must include the evaluators, decision-makers, stakeholders, and any other personnel familiar with the intended outcomes of the program (Fitzpatrick et al., 2004).

Product Evaluation

Product evaluation, which is the last element in the CIPP Evaluation Model, refers to the ultimate decision associated with the fate of the program (Fitzpatrick et al., 2004). This decision may include continuation, termination, modification, or refocusing of the program under review (Stufflebeam, 2003). The outcome is a product of collections of descriptions and numerous archived judgments about the objectives, merit, and worth of the program. In order for the decision maker to arrive at a conclusion, the evaluator must collect both quantitative and qualitative information from all personnel and stakeholders involved. If necessary, the product evaluation can be divided into subcategories of impact, effectiveness, sustainability, and transportability in order to gain more concise information about the long-term effects of the program.

Systematic Analysis of Multiple Perspectives

As developed by Stufflebeam (2003), the CIPP Evaluation Model is multifaceted, requiring a systematic analysis of multiple perspectives, implementation and collection of both qualitative and quantitative data, and the employment of triangulation procedures (Fitzpatrick et al., 2004). To collect a wide array of information that addresses a variety of opinions, the model encourages evaluation methods that focus on multiple perspectives. These perspectives are accounted for by assigning various stakeholders the task of qualitative data collection. Although data gathering techniques may be self-made or procedural, the intent is to gain valid information about the functionality of the program (Stufflebeam, 2003).

Exemplifying its flexibility, the model can be utilized for either formative or summative purposes or both. Consistent with its focus on improving the program, this model was designed primarily to facilitate planning and

implementation efforts. According to Stufflebeam (2003), this evaluative approach is meant to "supply evaluation users—such as policy boards, government officials, foundation presidents and staff members, project staffs, school administrators" with concrete information that will promote sound goals and forward thinking (p. 4). In addition, by examining retrospective, or summative uses, this model enables potential consumers to analyze the quality, cost-effectiveness, and utility of a potential program. Moreover, evidence can be produced to provide defensible information concerning cost, time, or effectiveness.

Addresses Single or Multitude of Needs

Used to address issues typically associated with accountability, educational, planning, research, and business purposes, the CIPP Evaluation Model is an evaluative measure directed at addressing either a single need or a multitude of needs (Stufflebeam, 2003). Consequently, this model can function in one of two ways: a non-linear, non-sequential evaluation approach; or a linear, sequential evaluation approach (Fitzpatrick et al., 2004). The consistency of this model is derived from its ability to achieve improvement by providing the decision-maker with documented, clear, and unambiguous information concerning program processes and results.

The CIPP Evaluation Model has been employed as an evaluation model throughout the United States and worldwide for both long- and short-term evaluative measures. Applications of the model have spanned numerous disciplines, including education, residential development, transportation safety, and governmental review systems.

Although Stufflebeam designed the CIPP Evaluation Model to provide flexibility, some believe that strict adherence to each component within the model is vital. Some evaluators become so committed to this particular evaluation model that they are convinced that skipping any of the four components will result in a shortage of information needed by decision-makers. According to Felix (1979), an educational evaluator, "It is vital that each model for school evaluation take all four CIPP stages into account" (p. 58). Felix does note, however, that the nonlinear design of this model can be used advantageously to meet the needs of a variety of contexts. Similarly, in an article entitled *Guidelines for Evaluating Parent Training Programs*, Matthews and Hudson (2001) emphasize that the components of this model ensure continuous evaluative procedures by accounting for both pre- and post-assessment.

On the other hand, when Frasier (2004) undertook responsibility for evaluating the *Wisconsin Special Education State Improvement Grant*, he employed an adaptation of the CIPP method in conjunction with other published evaluative procedures. This adaptation enabled the evaluation

to focus on the three main identified goals. Rather than strictly adhering to existing evaluative measures, Frasier was able to modify and adapt the model to account for the specific information needed. Similarly, an evaluation report concerning a "Healthy Students Initiative" utilized only two of the four parts of the CIPP Evaluation Model. Because this particular evaluation was not the first for the program, the issues of context and input had already been addressed (Lohmeier, Lee, Frey, Tollefson, & Bailey, 2003).

With the emergence of new governmental programs and educational incentives, this model has proven to be a beneficial means of conducting program evaluations to ensure accountability (Stufflebeam, 1971b). Accountability refers to "the ability to account for past actions in relationship to the decisions which precipitated the actions, the wisdom of those decisions, the extent to which they were adequately and efficiently implemented, and the value of their effects" (p. 20). Because the model focuses on improvement, many federal grants have utilized the CIPP framework as a means of assessing current programs and identifying areas needing improvement in subsequent programs (Fitzpatrick et al., 2004).

Though applicable governmental and federal reports, the CIPP Evaluation Model is widely utilized in educational evaluations (Stufflebeam, 2003). In a journal article relating to implementation of evaluative models in assessing reading instruction, Nicholson (1989) stated that this model is "a useful model for the classroom situation" (p. 313). Citing the importance of accountability, Nicholson goes on to outline various intended outcomes associated with the execution of this model. Moreover, exemplary outcomes of the model are illustrated through the discussion of methods concerning the probable application of needs-based assessment, varied instructional tools and strategies, and product implementation.

Continuing the focus of accountability, Stufflebeam (1971a) discussed the relationship of this model with educational accountability. Perhaps the most fundamental and accessible aspect of accountability is the analysis of archived data. Foremost in this model is identifying the context of a program evaluation to specify objectives. These specifications provide a record of intended goals and the purpose of their implementation. Identified objectives and intended outcomes, Stufflebeam noted, are the essential bases of documented accountability. Second, input evaluation, or the specific strategy utilized within a program, provides evidence of accountability by identifying any additional strategies under consideration. Process evaluation, the third step in this model, serves as written documentation of the actual processes that occurred. Therefore, accountability can be assessed on the combined bases of implementation and performance (Stufflebeam, 2003). Finally, the process of comparing outcomes with intended objectives, or product evaluation, also ensures program accountability.

While the CIPP Evaluation Model is widely utilized throughout the educational field, many university graduate programs still exhibit a preference for a more results-oriented approach. In 1996, a multi-collegiate evaluation was conducted to determine the status of management-oriented evaluation approaches in public administration and management graduate programs. Citing this model as an evaluation design that focuses on improvement, the findings of this report revealed that "courses are more likely to focus on program outcomes and policy impact than management issues" (Madison, 1996, p. 251). Madison maintained that administrative and management courses should increase their focus on teaching students to emphasize evaluations as effective management tools. Ultimately, research suggests that evaluation strategies are a venue to organizational improvement. She further emphasized that expanding the evaluation content in public administration programs would result in increased focus on management-oriented approaches.

Employing the Kirkpatrick Model

Kirkpatrick (1998) originally discussed an evaluation of a training program, which later developed into a four-step, hierarchical model. Although it has been used less as evaluations become more difficult and costly, Kirkpatrick's model remains the standard for business and industry (Allinger & Janak, 1989). In ascending order, the four levels of Kirkpatrick's model are *reaction, learning, behavior,* and *results.* Evaluation of trainees' reactions to training is the easiest and least costly step in Kirkpatrick's Model. It has been reported that up to 95% of training evaluations occur at this level (Arthur et al., 2002b).

Measuring Reaction of Learners

Reaction has been loosely defined as how favorably learners react to training, "learner satisfaction" as participant satisfaction, and as a measure of customer satisfaction (Arthur et al., 2002a; Kirkpatrick, 1998). According to Allinger and Janak (1989), the reaction that Kirkpatrick was referring to was how the trainees' liked or felt about the program and whether it was attitudinal (affective) rather than behavioral; thus, Kirkpatrick stressed attitudes, or the affective domain, of learning. Reaction refers to the trainees' like or dislike of a program (Boyle & Crosby, 1997). It provides feedback; it can signal that instructors care about trainees' satisfaction; it can provide concrete quantitative data for decision makers, managers, deans, academic vice-presidents; it can provide instructors with quantitative information that can be used to establish standards for future performance and learning

such as specific numerical goals for the average satisfaction of future participants (Schumann et al., 2001).

Data obtained from trainees at this level can be a collection of reactions to the instructor, to the course, and to the learning environment; the reactions are measured. Responses may be to the session or course, to the trainer's presentation or style, and/or to the content and quality of the training materials. Key evaluation techniques include asking, listening, or using evaluation forms at the conclusion of a course. Reaction can provide administrators and instructors with valuable insights for course improvement in areas where students' input is the best data. This could possibly be viewed as a form of formative evaluation (Boyle & Crosby, 1997).

Some of the evaluations or questionnaires used in the reaction level have been called happiness sheets, smile sheets, or happy sheets because in their simplest form they measure how well students like training. However, reaction level questionnaires can reveal valuable data if more complex questions are asked. Questions about the relevance of the objectives, ability of the course to maintain interest, amount and appropriateness of interactive exercises (online), ease of navigation, perceived value, and transferability to the workplace could all be submitted. Reaction can be used to assess formative evaluations (Birnbrauer, 1987; Kruse, n.d.). Level 2 involves the evaluation of trainees' learning and is a slightly more complicated and costly step in Kirkpatrick's model than Level 1.

Measuring Learning Due to Training

It has been reported that up to 37% of training evaluations occur at Level 2, which measures changes in *learning* due to training (Arthur et al., 2002b; Schumann et al., 2001). More specifically, Level 2 measures the extent to which participants experience changes in attitudes, skills, or motivations as a result of the training (McGivern & Bernthal, 2002). The changes measured by Level 2 can also be characterized as change in the *cognitive domain* (knowledge and factual information), *psychomotor domain* (skills), and *affective domain* (attitudes). The cognitive domain is commonly assessed with multiple choice, open-ended, and listing-of-facts-type tests that can be assessed either immediately or later to determine degree of knowledge retained. The psychomotor domain is a performance test of skills measured in the training environment rather than in the actual work environment. Tests of skills measured in the work environment are assessed at the next level of the model. The affective domain focuses on how participants feel or think about the training content, on the implications of a participant's motivation to use training, on confidence in using skills, on the ability to reach goals, on

the interpretation of information, facts, and principles, on the techniques understood and absorbed, and the degree to which the participants have learned (Allinger & Janak, 1989; Arthur et al., 2002b; Boyle & Crosby, 1997; Kirkpatrick, 1998; McGivern & Bernthal, 2002).

Data collected from trainees' at this level can use media such as text, voice, or demonstration. It is best to use testing measures both before and after training and to aim for a 100% response level (Kirkpatrick, 1998). Common tools of measurement at this level may include interviews, surveys, and tests (pre- and post) with control groups, the latter allowing for the best comparison of the change observed in the experimental group against the change observed in the control group. The best approach is to randomly assign students to one of the two groups. If it is evident that at the start the learners in the two groups are equal in attitude, knowledge, and skills, then a only a post-test control group design is appropriate. Also, observations or combinations of the aforementioned tools can be used. Arthur et al. (2002b) mentioned the use of case studies and quizzes as well. In schools this is measured more often with examinations, quizzes, mid-terms, projects, and/or portfolio assessments common in educational settings (Boyle & Crosby, 1997).

Measuring Changes in Behavior

Level 3 measures changes in behavior in the workplace due to the training that has taken place. Evaluation at this level is much more complicated and costly than at Levels 1 and 2 as indicated by the mere 13% of training evaluations occurring here. Behavior here is defined is a transfer of knowledge, skills, and/or attitudes to the real world. Level 3 measures achievement of performance objectives (difference between learning and application) and involves the use of learned principles and techniques on the job (Allinger & Janak, 1989). Behavior evaluation assesses participants' ability to: carry out tasks more effectively after the course, provide better care to clients, and be more knowledgeable or skillful in job performance (Arthur et al., 2002b). It is necessary to collect data to the degree in which learners are using what they learned from a training course on their jobs. The transfer that has occurred in the learners' behavior must be due to the training program: Are the newly acquired skills, knowledge, or attitudes being used in the everyday environment of the learner? Many trainers consider this to be the truest assessment of a program's effectiveness. When to evaluate, how often, and in what manner are all important questions requiring answers (Winfrey, 1999).

Behavior is the degree in which the learners are transferring the attitudes, knowledge, and skills taught in one class to subsequent classes and nonacademic settings, such as the workplace (Schumann et al., 2001). A

change in behavior requires: a desire to change, necessary knowledge and skills, and the "right" job climate. It might require encouragement, help, and rewards. Terms such as *preventing*, *discouraging*, *neutral*, *encouraging*, and *requiring* might be used in a Likert-type scale on a behavior level form measuring job performance (Kirkpatrick & Kirkpatrick, 2005). According to Kirkpatrick, performance can be improved on the job through appraisal and coaching.

Measuring Results of Training

Level 4 examines organizational results and is the most complicated and costly step in the model. It assesses "the bottom line" and the final results, but only 3% of training evaluations occur at this level (Arthur et al., 2002b). It identifies how training changes organizational functions such as grievances, turnover, reductions in costs, and absenteeism (Birnbrauer, 1987). The definition of "results" depends upon the goals of the training program. Goals might include ends, results desired, reduction of costs and turnover, reduction of absenteeism and grievances; increase in quality and quantity of production, or improving morale (Allinger & Janak, 1989). Without sufficient awareness of the actual results, a good academic program may be eliminated because there is no data to prove what it accomplished and what it offers the many stakeholders involved such as students, faculty, employers, parents, and the community (Boyle & Crosby, 1997). An organization needs at least some evidence, if not outright proof, of achievement at this level (Kirkpatrick, 1977b). Results measure the degree to which the output of the participant's workgroup or organization has improved because of the learning program (Kirkpatrick, 1998).

The results level is the most difficult level of the model to measure (McGivern & Berthal, 2002). The factors to be gauged will depend upon the field or practice in which the program under review is established. For example, measurements that might impact sales training could include sales volume, customer retention, length of sales cycles, and the profitability of each sale; measurements that might impact technical training could include a reduction in calls to the help desk, less time necessary to complete reports or tasks, and improved use of software or computer systems. A measurement that might impact quality training could be a reduction in the number of defects, while measurements that might impact safety training could include a reduction in the number or severity of accidents, and a measurement that might impact management training could be an increase in engagement levels of direct-reports (Kruse, n.d.). It is important to decide which results are relevant to the evaluation and decision-making process. In a similar way, job parameters for workers or trainees could include salary increases, more job offers, better promotions, higher productivity, higher

quality work, lower costs, fewer accidents, higher sales, lower turnover, or higher profits (Schumann et al., 2001).

Level 4 evaluations depend upon the objectives of the training program and must be quantified. The evaluator must determine the level of proof or evidence required by the stakeholders (Kirkpatrick, 1977b). Kirkpatrick points out the differences between evidence and proof at each level of the model. Although evidence can be seen as similar to that needed to convict a person in a courtroom, proof is very difficult, or even impossible, to come by. Evidence in a courtroom is not absolute proof, but it must be fairly strong in order to implicate guilt. During a Level 1 evaluation, proof can be obtained through honest answers by making responses anonymous. At Levels 2 and 3 it is more difficult to obtain proof but, through the use of control groups and the proper tests, one can arrive at some degree of confirmation It is so difficult to establish proof at Level 4 of Kirkpatrick's model that he has proposed that one must either "provide proof at all costs or convince the boss that evidence is good enough and that proof is either impossible, or at least impractical" (p. 13). Evidence that one's training is effective is less expensive and easier to come by than absolute proof. Absolute proof of either changed behavior or results impacting the organization due to training is probably unattainable, or attainable only through a scientific experimental design that is in all likelihood too expensive.

Measurable impacts on the organization could include cost versus benefit, proof versus evidence on report productivity, customer satisfaction, efficiency, morale, and profitability (McGivern & Bernthal, 2002). Results should be tracked over time, month to month, or before and after training (longitudinal study). This type of evaluation is expensive, time-consuming, and a drain on resources. Consequently, a decision to evaluate at Level 4 must include a cost analysis (McGivern & Bernthal) or *return on investment* (ROI). The evaluator must allow time for results to occur (6-12 months) instead of assuming instantaneous results. After the 6-to-12 month period has passed, the evaluator should repeat measurements of results at appropriate times. The use of a control group can always make the data more provable and allow for better statistical analysis. Kirkpatrick and L'Allier (2004) have provided useful guidelines to aid the evaluator's decision-making process (Table 2). The guidelines apply to all levels, not just Level 4.

Return of Investment

Return on Investment (ROI) is usually viewed as distinct from Level 4 because it is focused on a calculated percentage return. Gains observed from training intervention such as results or behaviors are converted into monetary values. They are then compared to the per person cost of the

Table 2

Evaluation Guidelines for Kirkpatrick Levels 1 to 4

Guideline	Level 1: Reaction	Level 2: Learning	Level 3: Behavior	Level 4: Results
Determine what you want to find out.	x			x
Design a form to quantify results.	x			
Encourage written comments.	x			
Get 100% response.	x	x		
Get honest responses using an anonymous instrument.	x			
Get delayed reaction.	x		x	x
Develop an acceptable standard or benchmark	x			
Measure against an agreed-upon standard	x			
Measure pre- and post-instruction.		x	x	x
Use paper/pencil for evaluating knowledge and attitudes.		x		
Use performance measures for evaluating skills.		x		
Allow time for behavior to take place (opportunity).			x	x
Use a control group, if possible.		x	x	x
Survey a 360 degree observa-tion of new behaviors (by self, manager, subordinates, and others).			x	
Repeat evaluation at appropriate times.			x	x
Determine the cost of evaluation versus benefits.			x	x

Source: Kirkpatrick & L' Allier, 2004, p. 32

training (McGivern & Bernthal, 2002). The cost of training and development must be compared to the ROI to decide whether or not a particular training or evaluation should be continued. The ROI can also be used to help decide whether changes need to be made in evaluation techniques or in training, and thus it can be useful during formative evaluation. The return on investment is sometimes difficult to measure and explain and can be more trouble than it is worth to ascertain. ROI has become increasingly

important to investors in training programs as global competition has increased and fiscal belts have tightened. It refers to the return that the training program, including the evaluation of the training, can have for the investor. It is used to indicate how training can be improved and whether it is necessary to remove a training program or parts of a program. The ROI can also indicate if the evaluation is "worth" the investment. It is a method of holding trainers and evaluators accountable to the investors (Brauchle & Schmidt, 2004; Kirkpatrick & Kirkpatrick, 2005; Riotto, 2004).

The fifth level was added because the Kirkpatrick model only focused upon the effectiveness of training events and not the monetary benefits. Instructional developers needed a way to demonstrate cost-value and accountability for their training. The original idea was published by Jack Phillips to quantify the monetary value of training investments. ROI answers the question, "For every dollar invested in training, how many dollars does the employer get back, in other words, show me the money?" Some researchers have mentioned that the ROI converts the fourth level (business results) into monetary values and compares it with the training costs, the only monetary measure of the value of the training. This provides information to improve the training (Arthur et al., 2002a). Phillips claimed that "many programs have failed to deliver what was expected; consequently, program sponsors have requested ROI calculations/justifications," and, "the concern for accountability in all functions in organizations is increasing: Thus, the training and development function becomes one of many support efforts under scrutiny" (p. 3).

The process involved in Kirkpatrick's model is similar to that of other evaluations. Although it is often understood to be a hierarchical model, Kirkpatrick's model could be better viewed as a group of potential evaluation tools or, as some have suggested, taxonomies of evaluation (Holton, 1996a; Kirkpatrick, 1996; Kirkpatrick & L'Allier, 2004). Viewed in this manner, the model actually becomes a decision-making process that takes into account the questions that must be answered and then applies the level or levels from the model that need to be utilized for an effective evaluation. The evaluator must first consult with stakeholders to determine the objectives and goals of the training. The team must then decide which questions can possibly be answered by training. Next, the evaluator, through appropriate consultation, will decide upon the appropriate level or levels of the Kirkpatrick model for the training evaluation. Although most evaluations of training have been at the reaction level, in reality an organization probably does desire to know the "bottom line" effect of training. Therefore, it is likely that at least some evaluation at Level 4 will be desired by the investors. Generally, interest goes beyond what the trainees "feel" about the training. Training should involve actual learning, and learning should translate or "transfer" to positive behavioral changes

in the workplace, followed by organizational improvements. Ideally, all four steps of the model should be applied. School settings already apply Level 2 when they give mid-term exams, quizzes, and other types of tests.

The evaluator must decide what questions are to be answered, which level of evaluation to apply, and, along with the stakeholders, the cost of the evaluations and/or training. Once all of these are considered, the evaluator will develop the proper tools (assessment forms) and strategies to be used at the levels decided upon. Once data are collected and compiled, statistical analyses will be applied and then reports for stakeholders developed. According to Birnbrauer (1987), depending upon the questions asked, additional evaluation points may be needed, such as Level 5 (ROI) or an evaluation matrix for educational-type evaluations.

Outcome-Base Evaluation Processes

In an Outcome-Base Evaluation (OBE), four types of evaluation are most commonly used. The first is *program evaluation.* This is the type of evaluation that uses either person-referenced or organization-referenced outcomes to determine whether the program is meeting the desired outcomes and uses. The next type is *effectiveness evaluation*, which strives to report the extent to which a program is meeting its goals and objectives. Another method is *impact evaluation*. This evaluation studies whether or not a program has made a difference for its stakeholders compared to an alternative program. Finally, *policy evaluation* researches the equity, efficiency, or effectiveness of policy outcomes for a program (Schalock, 2001).

Program Evaluation

Program evaluation typically strives to answer the question, "What outcome is my program producing in its service recipients?" Again, these evaluations are used most often in the education, health, and social services realms. Program outcomes generally fall into four categories: organization performance outcomes, organization value outcomes, individual performance outcomes, and individual value outcomes. Each possible outcome is aggregated into a cell in the matrix where it is most likely to fall. Organizational outcomes deal with staff, staff competence, and customer satisfaction with the program. Individual outcomes are more related to personal preferences, qualities, and values.

A program evaluation generally involves three steps. First, the program must establish a *baseline* established by the outcomes that are currently being used. To establish a baseline, organizations would fill out a detailed

inventory, and the evaluators would code it and determine the emphasis of their evaluation. The next step involves ascertaining the *desired outcomes.* This tells the evaluator the direction of the organization. Usually, the organization fills out a second inventory, this time choosing the responses that indicate desired goals. This will require focusing more on organization and individual value outcomes than was reflected in the organization's baseline. Finally, the evaluator will strive to explain how the organization can align its services with the desired outcomes.

The outcomes of a program must be measured in order to gauge the quality and effectiveness of education, health care, and social services. Most of these are measured through the perception of accountability. The outcomes must be measured for performance and consumer appraisal in the area of satisfaction, and also for functionality in the areas of adaptive behaviors and role status In addition, it will be looked at in the area of personal appraisal, which addresses life experiences and the quality of life.

In order for a program evaluation to be successful, an organization must be willing to foster a culture of change, be accountable in developing strategic plans and performance goals, and develop a monitoring system (Mika, 1996). Evaluations are more successful if stakeholders participate in the evaluation and are engaged in the decision-making process. Organizations that typically use program evaluations to their fullest potential have a well-defined authority within the organization, allocate sufficient time and resources to implementation, and are driven by their goals rather than by rules and regulations.

There are some *limitations* to program evaluation (Mika, 1996). Measured outcomes must be chosen wisely and balanced with other considerations. It would be too time consuming to measure all outcomes. That type of evaluation can be costly and an evaluator must be careful not to consume more resources during the evaluation than its information is worth. Third, the usefulness of this method varies. They are only useful to the management and decision-making personnel to the extent that they are used and that they answer the right questions. Finally, organizations are limited in their ability to influence or change some outcomes. Sometimes the organization is not the sole decision maker for an outcome. Many times it is beyond the control of the organization and is influenced by economic trends, demographics, and other factors.

Effectiveness Evaluation

Effectiveness evaluation strives to answer the question, "Is my program meeting its goals and objectives?" Its primary uses are to: (1) compare the program's goals with its achieved outcomes; (2) report the program's performance and value outcomes; and (3) provide formative

feedback information for program change and improvement (Schalock, 2001). This type of evaluation is very similar to program evaluation except effectiveness evaluation establishes a comparison condition against which accountability and outcomes can be judged. This model summarizes the five effectiveness analysis steps: (a) performance goals, (b) purpose and comparison condition, (c) methodology, (d) data collection and analysis, (e) and person- and organization-referenced outcomes. This model also involves a comparison group against which the significance of the results is compared.

To begin an effectiveness evaluation, the organization defines its strategic and performance plans and, from the findings, specifies the anticipated outcomes. Following this, the organization defines its purpose along with its comparison condition with which to compare significance. It is at this point that the organization specifies its outcome categories and measurement techniques. Outcome-based data should then be collected and analyzed. Following this step, the comparison can be done and outcomes can be determined and put into play. Using this model allows an evaluator to approach effectiveness evaluation systematically. Effectiveness evaluations can provide organizations with reliable data about their programs with an opportunity for recommendations for improvements (Mika, 1996). The evaluation process will assist policy makers' responses in determining issues such as, Are objectives and outcomes being met? and, Are resource values being maintained?

However, implementing an effectiveness evaluation that will provide the information required to continuously improve organizations will not be an easy task. Some of the challenges facing evaluators include: (a) development of policy, objectives, and terms of reference; (b) establishment of a monitoring framework; (c) location of resources ; (d) management, analysis, and interpretation of data, and presentation of results; (e) establishment of cooperative, long-term commitments between comparative organizations and other stakeholders; and (f) recommendation for changes in management styles within the organization from emphasizing prescriptive approaches and standardized rules to monitoring results and adapting actions.

Impact Evaluation

Impact evaluation stems from the question "Is my program better than others?" This type of evaluation focuses on the program's impacts and determines whether those impacts can be attributed to the intervention being employed or the services being evaluated. It also gives stakeholders

feedback in the areas of accountability and improvement plans (Schalock, 2001). Conducting an impact evaluation requires a comparison group with which to compare results or outcomes. The evaluator must look at the people served by each program, the services rendered by each, the outcomes, and then determine whether there is a statistically significant difference in the results.

There are six evaluation designs most commonly used in impact evaluations: (1) experimental/control, (2) matched pairs, (3) hypothetical comparison group, (4) longitudinal status and comparison, (5) pre- and post-change comparison, and (6) person as own comparison (Schalock, 2001). There is a direct relationship between design choice and certainty, precision, comparability, and generalizability of the results. As with any research, the closer an evaluator can come to a true experimental design, the better the results in the aforementioned categories. However, this type of design may not always be feasible since this discussion focuses on human beings in the areas of education, health care, and social services. As a cautionary note, it must be mentioned that an evaluator should not force him or herself to use an experimental design unless he or she can randomly assign participants into different groups for comparison. The other designs are reasonable alternatives, but the evaluator must be aware of their shortcomings.

When choosing an evaluation design, the evaluator must know the purpose for the data and the evaluation, which is to determine if one program has made a difference compared to another program (Mika, 1996). The evaluator must establish the comparison condition, describe the data sets and how they were measured, and present the results. These results include making an impact statement based on significant mean differences. Finally, the evaluator will discuss the major results and their implications for the program. According to Priest (2001), it is then up to the program to make the suggested changes or policy improvements.

As one type of OBE, an impact evaluation's results may appear to be less applicable for continuous improvements. However, if an evaluator is successful in identifying factors that account for differences between programs and variables that produce successful programs, then the results can be very beneficial to that organization. Another positive aspect is that often impact evaluations provide the framework for the next type of OBE policy evaluation.

Policy Evaluation

Policy evaluation strives to answer the simple question, "Does this policy work?" This method of evaluation determines outcomes in reference to their equity, efficiency, and effectiveness. Policy analysis should focus on the same standards as other OBEs: performance and values (Schalock,

2001). This type of evaluation employs multiple methods and can focus on the individual, the program, or the system as a whole. Also, policy evaluation techniques can vary and include any of the previously discussed evaluations, as well as a cost-benefit analysis. A cost-benefit analysis merely determines if a program's benefits outweigh its costs (Priest, 2001).

Policy evaluation involves five steps. First, the evaluator must describe the policy goals, the intent of the analysis, and its content. He or she must identify the values that underlie the policy. Then the evaluator must analyze the intended outcomes in reference to the stated goals. The focus should be on the interaction between the policy and the groups the program impacts. The goals are then fit into their respective cells by identifying key factors, and the status of the anticipated outcomes must be evaluated. Finally, the evaluator will provide feedback to the key evaluation players.

When performing a policy evaluation, the evaluator must keep in mind the validity, the importance, the usefulness, and the feasibility of his findings and also of the recommendations. Policy evaluation does not occur in a vacuum. There are a number of stakeholders who need to be involved, including the consumer, the practitioner, the supervisors and managers, and the policymakers who set the standards (Mika, 1996). An evaluator must keep in mind that it is easy to exaggerate the importance of government policies. Also, this type of evaluation can offer solutions to problems when there is no agreement on the nature of the problems. Policy evaluation deals with subjective topics and must rely on evaluators to interpret data. Policies are intended to produce outcomes; policy evaluation helps stakeholders capture and report those outcomes from many different perspectives.

Current education, health care, and social service programs are increasingly being asked to demonstrate their effectiveness and efficiency (Eliot, 2003; Johnson, 2000; Shapiro & Derrington, 2004). This is impacting all organizations, and they are responding at the program level by changing the way they do business and conduct program evaluations. They are using outcome-based data as a basis for making changes to improve services and increase accountability.

During the past several years there has been a shift in educational evaluations from process to outcomes. Policy makers are looking at outcomes such as graduation rates, student test scores, teacher proficiency rates, accreditation, parent involvement, school climate, student satisfaction, academic skills, social adjustment, social belonging, and empowerment (Kovalik & Dalton, 1998; Meehan, Walsh, Swisher, Spring, & Lewis, 1999; Shapiro & Derrington, 2004; Ziegler & Sussman, 1996). Health care evaluations also have shifted their focus to look at outcomes that reflect how well individuals can function in daily life, as well as their perceived well-being (Ciarlo, 1982; Eliot, 2003; Hargreaves, 1982). Health care

evaluators study such outcomes as clinical symptoms, coping behaviors, feelings, best care, costs of health care, advantageous gains and balance, mortality, longevity, and customer satisfaction. Social services are another area in which outcome-based evaluations are heavily used. The outcomes typically assessed in this area are measures such as optimality, public engagement, consumer satisfaction, accessibility to services, waiting lists, employment status, social inclusion, legal status, education status, material well-being, self determination, and rights (Mika, 1996).

Advantages and Disadvantages

No evaluation model can be all things to all people in all situations. Kirkpatrick's model has stood the test of time since it was first outlined in 1959, but not without controversy. Critics charge the reactions and learning assessment identified in Levels 1 and 2 are of limited value in determining the effectiveness of training and supply little or no explanation of how learning has occurred. In fact, some in the field insist the data cannot be validated because there is little correlation between how trainees feel and what is actually learned (Birnbrauer, 1987). Levels 3 and 4 are considered costly and are often omitted from the evaluation process as a result. However, because the Kirkpatrick model is so flexible, each level can be administered separately from the rest. Where quantitative measurements are best, Kirkpatrick's model has shown widespread usefulness and fits well into management-oriented environments. Quantitative measurement forms contained in Kirkpatrick's model can easily be adapted to the situational need of the evaluation.

The CIPP Model has gained in popularity in recent years. CIPP focuses on improving processes and was designed for use in the planning and implementation stages. It is also useful in assessing cost and utility in retrospect to determine the effectiveness of a program. Although its author intended to provide flexibility, some evaluators believe that strict adherence to each component is necessary. However, other literature reports successful use of modified versions of the CIPP for specific information needs. In an era of accountability, especially in governmental programs, the CIPP Model provides a means for assessing programs and identifying areas for improvement.

Outcome-based evaluation (OBE) has gained in popularity with the changing times. In the recent past there has been a shift from process to outcome in the areas of education, health care, and social services. Outcome-based evaluation can be very beneficial to the client and is relatively easy to adapt depending upon the desired outcome. It can also improve a program almost immediately if the stakeholders are open to change and willing to make program improvements.

Conclusion

Program evaluation is here to stay. Tightening budgets and increased competition have become the norm in both governmental arenas and private enterprise. Evaluating program efficiency, effectiveness, and success will be the key ingredient in making adjustments that are both timely and cost effective. Evaluation models that provide practical tools and easily understood directions leading to verifiable results in a cost-efficient manner will be of the utmost importance. For any program to continue operating in this era of high-pressure, high-stakes business, governmental, educational, and individual endeavor, it must consistently and efficiently meet the needs of its clients. Those who serve must be in a state of constant self-examination in order to stay vital. It is a buyer's market, so let the seller beware.

References

Allinger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personal Psychology, 42,* 331-342.

Arnold, E. (2003, December 1). How can businesses measure the impact of training? *Mass High Tech: The Journal of New England Technology.* Retrieved March 18, 2006, from http://masshightech.bizjournals.com/masshightech/stories/ 2003/12/01/focus6.html

Arthur, K., Gage-Little, R., & Munson, D. (2002a). *Evaluating training programs level 5: Return on Investment.* Retrieved March 21, 2006, from http://www.dalemunson.com/ products/ Kirkpatrick2Jack%20PhillipsROI.pps

Arthur, K., Gage-Little, R., & Munson, D. (2002b). *Training Kirkpatrick's four levels of evaluation.* Retrieved March 21, 2006, from http://www.dalemunson.com/ products/Kirkpatrick%20Presentation.pps

Baskin, C., Barker, M., & Woods, P. (2005). When group work leaves the classroom does group skills development also go out the window? *British Journal of Educational Technology, 36*(1), 19-31.

Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning, 27,* 341-347.

Birnbrauer, H. (1987). Evaluation techniques that work. *Training and Development, 41*(7), 53-55.

Boyle, M. A., & Crosby, R. (1997). Academic program evaluation: Lessons from business and industry. *Journal of Industrial Teacher Education, 34*(3), 81-85.

Brauchle, F., & Schmidt, K. (2004). Contemporary approaches for assessing outcomes on training, education, and HRD programs. *Journal of Industrial Teacher Education, 41*(3), 1-12.

Chambers, F. (1994). Removing confusion about formative and summative

evaluation: Purpose versus time. *Evaluation and Program Planning, 17,* 9-12.

Chen, H., & Rossi, P. H. (1989). Issues in the theory-driven perspectives. *Evaluation and Program Planning, 12,* 299-306.

Ciarlo, J. A. (1982). Accountability revisited: The arrival of client outcome evaluation. *Evaluation and Program Planning, 5,* 31-36.

Eliot, S. (2003). *Healthy Child Care Colorado, 2002: Outcome evaluation.* Denver, CO: Colorado Department of Human Services, Denver Division of Child Care. (ERIC Document Reproduction Service No. ED 479 260)

Felix, J. (1979). Research and evaluation to improve instruction: The Cincinnati strategy. *Educational Evaluation and Policy Analysis, 1*(2), 57-62.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston: Pearson.

Frasier, J. R. (2004). *Wisconsin Special Education State Improvement Grant (SIG).* Retrieved April 1, 2006, from http://www.wisconsinsig.org/ e&a/ SIG_Eval_Conceptual_Framework.pdf

Hargreaves, W. A. (1982). Outcome evaluation or treatment research? A response to Ciarlo. *Evaluation and Program Planning, 5,* 357-358.

Holton, E. F. III (1996a). The flawed four-level evaluation model. *Human Resource Development Quarterly, 7,* 5-21.

Holton, E. F. III (1996b). Final word: Response to "Reaction to Holton Article". *Human Resource Development Quarterly, 7,* 27-29.

Johnson, L. R. (2000). In service training to facilitate inclusion: An outcomes evaluation. *Reading and Writing Quarterly: Overcoming Learning Difficulties, 3*(16), 281-287.

Kirkpatrick, D. L. (1977b). Determining training needs: Four simple and effective approaches. *Training and Development, 31*(11), 22-15.

Kirkpatrick, D. L. (1996). Invited reaction: Reaction to the Holton article. *Human Resource Development Quarterly, 7,* 23-25.

Kirkpatrick, D. L. (1998). *Evaluating training programs: The four levels.* San Francisco: Berrett-Koehler.

Kirkpatrick, D. L., & L'Allier, J. J. (2004, June). *Evaluation as a strategic tool.* Chief Learning Officer, 30-33.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2005). *Transferring learning to behavior: Using the four levels to improve performance.* San Francisco: Berrett-Koehler.

Kovalik, C. L., & Dalton, D. W. (1998). The process/outcome evaluation model: A conceptual framework for assessment. *Journal of Educational Technology Systems, 27*(3), 183-194.

Kruse, K. (n.d.). *Beyond Kirkpatrick: Measuring the financial returns of e-Learning.* E-Learning Guru. Retrieved February 20, 2006, from http://www.e-learningguru.com/ articles/ art5 1.htm

Lohmeier, J. H., Lee, S. W., Frey, B. B., Tollefson, N., & Bailey, J. (2003).

Evaluation Report for the Safe Schools Healthy Students Initiative, Year 1: Lawrence Public Schools. University of Kansas, Center for Research, Inc. Retrieved April 1, 2006, from http://www.ku.edu/~spear/Documents/ SSHS/SSHS%20Evaluation%20Report%20Year%201.pdf

Madison, A. M. (1996). The status of management oriented evaluation in public administration and management graduate programs. *Evaluation Practice, 17*(3), 251-259.

Matthews, J. M., & Hudson, A. M. (2001). Guidelines for evaluating parent training programs. *Family Relations, 50*(1), 77-86.

McGivern, M. H., & Bernthal, P. (2002, Winter). Measuring training impact. *The Catalyst.* Retrieved March 28, 2006, from http://www.findarticles.com/ p/articles/mi_qa4011/is_200201/ai_n9057654.

McNamara, C. (2000). *The field guide to nonprofit program design, marketing, and evaluation.* Minneapolis: Authenticity Consulting. Retrieved March 25, 2006, from http://www.mapnp.org/library/evaluatn/fnl_eval.htm

Meehan, M. L., Walsh, S., Swisher, A., Spring, J., & Lewis, H. (1999, April). *Process and outcome evaluation of an Even Start program.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Mika, K. L. (1996). *Program outcome evaluation: A step-by-step handbook.* Milwaukee, WI: Families International.

Newstrom, J. W. (1978). Catch-22: The problems of incomplete evaluation of training. *Training & Development, 32*(11), 22-24.

Nicholson, T. (1989). Using the CIPP model to evaluate reading instruction. *Journal of Reading, 32*(4), 312-318.

Nickols, F. (2004). A stakeholder approach to evaluation training. *Distance Consulting.* Retrieved February 23, 2006, from http://www.nickols.us (cited with permission of Fred Nickols via email).

OMB: The Executive Office of the President. (1993). *Government performance results act of 1993* [On-line]. Available: http://www.whitehouse.gov/omb/mgmt-gpra/gplaw2m.html

Posavac, E. J., & Carey, R. G. (2007). *Program evaluation: Methods and case studies* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Priest, S. (2001). A program evaluation primer. *Journal of Experiential Education, 1*(24), 34-40.

Riotto, J. J. (2004). Model for calculating ROI of training/learning initiatives. *Journal of Interactive Instruction Development, 16*(4), 18-21.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed). Newbury Park: CA: Sage.

Royse, D., Thyer, B. A., Padgett, D. K., & Logan, T. K. (2006). *Program evaluation: An introduction* (4th ed.). Belmon, CA: Thomson and Brooks/Cole.

Sanders, J. R., & Sullins, C. D. (2006). *Evaluating school programs.* Thousand Oaks, CA: Corwin Press.

Schalock, R. L. (2001). *Outcome-based evaluation* (2nd ed.). New York:

Kluwer Academic/Plenum Publishers.

Schumann, P. L., Anderson, P. H., Scott, T. W., & Lawton, L. (2001). A framework for evaluating simulations as educational tools. *Developments in Business Simulation and Experiential Learning, 28,* 215-220.

Scriven, M. (1999). *The logic and methodology of evaluation: A monograph in the Sage Applied Social Research Series.* Retrieved March 25, 2006, from http://www.wval.cgu.edu/ lectures/intro/lomev1.htm

Shapiro, B. J., & Derrington, T. M. (2004). Equity and disparity in access to services: An outcome-based evaluation of early intervention child find in Hawaii. *Topics in Early Childhood Special Education, 24*(94), 199-212.

Stufflebeam, D. L. (1966). A depth study of the evaluation requirement. *Theory Into Practice, 6,* 126-133.

Stufflebeam, D. L. (1971a). The use of experimental design in educational evaluation. *Educational Measurement, 8*(4), 267-274.

Stufflebeam, D. L. (1971b). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education, 5*(1), 19-25.

Stufflebeam, D. L. (2003). *The CIPP model for evaluation.* Portland, Oregon: Annual Conference of the Oregon Program Evaluators Network (OPEN). Retrieved March 25, 2006, from http://www.wmich.edu/evalctr/pubs/CIPP-ModelOregan10-03.pdf

The United Way of America. (1996). *Measuring program outcomes: A practical approach.* Alexandria, Virginia: Author.

Watkins, R., Leigh, D., Foshay, R., & Kaufman, R. (1998). Kirkpatrick plus: Evaluation and continuous improvement with a community focus. *Educational Technology Research and Development, 46,* 90-96.

Winfrey, E. C. (1999). Kirkpatrick's four levels of evaluation. In B. Hoffman (Ed.), *Encyclopedia of Educational Technology.* Retrieved March 23, 2006, from http://coe.sdsu.edu/eet/articles/k4levels/starrt.htm

Ziegler, S., & Sussman, S. B. (1996). *The effectiveness of adult literacy education: A review of issues and literature related to outcome-based evaluation of literacy programs.* Toronto, Canada: Ontario Literacy Coalition. (ERIC Document Reproduction Service No. ED 399 423)

Dr. Ernest W. Brewer is a Professor and Principal Investigator/Director of Pre-College Enrichment Programs at The University of Tennessee. Dr. Brewer teaches a graduate course entitled "Program Evaluation in Education." His co-authors are *Elizabeth Peters, John Cummins,* and *Heather Kindall.* They were former doctoral students who took his Program Evaluation class.