2012; 34: e288–e299 MEDICAL TEACHER

WEB PAPER
AMEE GUIDE

Program evaluation models and related theories: AMEE Guide No. 67

ANN W. FRYE[1] & PAUL A. HEMMER[2]

[1]Office of Educational Development, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555-0408, USA, [2]Department of Medicine, Uniformed Services, University of the Health Sciences, F. Edward Hebert School of Medicine, Bethesda, MD, USA

Abstract

This Guide reviews theories of science that have influenced the development of common educational evaluation models. Educators can be more confident when choosing an appropriate evaluation model if they first consider the model's theoretical basis against their program's complexity and their own evaluation needs. Reductionism, system theory, and (most recently) complexity theory have inspired the development of models commonly applied in evaluation studies today. This Guide describes experimental and quasi-experimental models, Kirkpatrick's four-level model, the Logic Model, and the CIPP (Context/Input/Process/Product) model in the context of the theories that influenced their development and that limit or support their ability to do what educators need. The goal of this Guide is for educators to become more competent and confident in being able to design educational program evaluations that support intentional program improvement while adequately documenting or describing the changes and outcomes—intended and unintended—associated with their programs.

Introduction

Program evaluation is an essential responsibility for anyone overseeing a medical education program. A "program" may be as small as an individual class session, a course, or a clerkship rotation in medical school or it may be as large as the whole of an educational program. The "program" might be situated in a medical school, during postgraduate training, or throughout continuing professional development. All such programs deserve a strong evaluation plan. Several detailed and well written articles, guides, and textbooks about educational program evaluation provide overviews and focus on the "how to" of program evaluation (Woodward 2002; Goldie 2006; Musick 2006; Durning et al. 2007; Frechtling 2007; Stufflebeam & Shinkfield 2007; Hawkins & Holmboe 2008; Cook 2010; Durning & Hemmer 2010; Patton 2011). Medical educators should be familiar with these and have some of them available as resources.

This Guide will be most helpful for medical educators who wish to familiarize themselves with the theoretical bases for common program evaluation approaches so that they can make informed evaluation choices. Educators engaged in program development or examining an existing educational program will find that understanding theoretical principles related to common evaluation models will help them be more creative and effective evaluators. Similar gains will apply when an education manager engages an external evaluator or is helping to evaluate someone else's program. Our hope is that this Guide's focus on several key educational evaluation models

Practice points

- Educational programs are fundamentally about change; program evaluation should be designed to determine whether change has occurred.
- Change can be intended or unintended; program evaluation should examine for both.
- Program evaluation studies have been strongly influenced by reductionist theory, which attempts to isolate individual program components to determine associations with outcomes.
- Educational programs are complex, with multiple interactions among participants and the environment, such that system theory or complexity theory may be better suited to informing program evaluation.
- The association between program elements and outcomes may be non-linear—small changes in program elements may lead to large changes in outcomes and vice-versa.
- Always keep an open mind—if you believe you can predict the outcome of an educational program, you may be limiting yourself to an incomplete view of your program.
- Choose a program evaluation model that allows you to examine for change in your program and one that embraces the complexity of the educational process.

Correspondence: Ann W. Frye, Office of Educational Development, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555-0408, USA. Tel: 409-772-2791; fax: 409-772-6339; email: awfrye@utmb.edu

ISSN 0142–159X print/ISSN 1466–187X online/12/050288–12 © 2012 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.668637

in the context of their related theories will enrich all educators' work.

A focus on change

We believe that educational programs are fundamentally about change. Most persons participating in educational programs—including learners, teachers, administrators, other health professionals, and a variety of internal and external stakeholders—do so because they are interested in change. While a program's focus on change is perhaps most evident for learners, everyone else involved with that program also participates in change. Therefore, effective program evaluation should focus, at least in part, on change: Is change occurring? What is the nature of the change? Is the change deemed "successful"? This focus directs that program evaluation should look for both intended and unintended changes associated with the program. An educational program itself is rarely static, so an evaluation plan must be designed to feed information back to guide the program's continuing development. In that way, the program evaluation becomes an integral part of the educational change process.

In the past, educational program evaluation practices often assumed a simple linear (cause-effect) perspective when assessing program elements and outcomes. More recent evaluation scholarship describes educational programs as complex systems with nonlinear relationships between their elements and program-related changes. Program evaluation practices now being advocated account for that complexity. We hope that this Guide will help readers: (1) become aware of how best to study the complex change processes inherent in any educational program, and (2) understand how appreciating program complexity and focusing on change-related outcomes in their evaluation processes will strengthen their work.

In this Guide, we first briefly define program evaluation, discuss reasons for conducting educational program evaluation, and outline some theoretical bases for evaluation models. We then focus on several commonly used program evaluation models in the context of those theoretical bases. In doing so, we describe each selected model, provide sample evaluation questions typically associated with the model, and then discuss what that model can and cannot do for those who use it. We recommend that educators first identify the theories they find most relevant to their situation and, with that in mind, then choose the evaluation model that best fits their needs. They can then establish the evaluation questions appropriate for evaluating the educational program and choose the data-collection processes that fit their questions.

Program evaluation defined

At the most fundamental level, evaluation involves making a *value* judgment about information that one has available (Cook 2010; Durning & Hemmer 2010). Thus educational program evaluation uses information to make a decision about the value or worth of an educational program (Cook 2010). More formally defined, the process of educational program evaluation is the "*systematic collection and analysis of information related to the design, implementation, and outcomes of a program, for the purpose of monitoring and improving the quality and effectiveness of the program.*" (ACGME 2010a). As is clear in this definition, program evaluation is about *understanding* the program through a routine, systematic, deliberate gathering of information to uncover and/or identify what contributes to the "success" of the program and what *actions* need to be taken in order to address the findings of the evaluation process (Durning & Hemmer 2010). In other words, program evaluation tries to identify the sources of variation in program outcomes both from within and outside the program, while determining whether these sources of variation or even the outcome itself are *desirable or undesirable*. The model used to define the evaluation process shapes that work.

Information necessary for program evaluation is typically gathered through measurement processes. Choices of specific measurement tools, strategies, or assessments for program evaluation processes are guided by many factors, including the specific evaluation questions that define the desired understanding of the program's success or shortcomings. In this Guide, we define "assessments" as measurements (assessment = assay) or the strategies chosen to gather information needed to make a judgment. In many medical education programs data from trainee assessments are important to the program evaluation process. There are, however, many more assessments (measurements) that may be necessary for the evaluation process, and they may come from a variety of sources in addition to trainee performance data. *Evaluation*, as noted earlier, is about reviewing, analyzing, and judging the importance or value of the information gathered by all these assessments.

Reasons for program evaluation

Educators often have both internal and external reasons for evaluating their programs. Primary external reasons are often found in requirements of medical education accreditation organizations (ACGME 2010b; LCME 2010), funding sources that support educational innovation, and other groups or persons to whom educators are accountable. A strong program evaluation process supports accountability while allowing educators to gain useful knowledge about their program and sustain ongoing program development. (Goldie 2006).

Evaluation models have not always supported such a range of needs. For many years evaluation experts focused on simply measuring program outcomes (Patton 2011). Many time-honored evaluation models remain available for that limited but important purpose. Newer evaluation models support learning about the dynamic processes within the programs, allowing an additional focus on program improvement (Stufflebeam & Shinkfield 2007; Patton 2011). After we describe some of the theoretical constructs that have informed both older and newer evaluation approaches, we will describe the older quasi-experimental evaluation model and then some of the newer, more powerful, models that are informed by more recent theories. We have selected evaluation approaches commonly used in medical education that illustrate the several theoretical foundations, but there are other useful approaches

that we could not include in this limited space. The list of recommended readings at the end of this Guide will direct interested readers to information about other evaluation approaches.

Theories that inform educational program evaluation models

We now consider theories relevant to evaluation models to set the stage for descriptions of common or useful evaluation models. Educational evaluation models were not developed with education theories in mind; rather, the theories that informed thinking about science and knowledge in general underpinned the development of evaluation models. We will therefore take somewhat of a historical approach to describing some of those theories and their relationship to the thinking of evaluation experts over the years. These same theories can now inform current educators' choices of evaluation models.

Reductionism

Many of the commonly used approaches to educational evaluation have their roots in the Enlightenment, when understanding of the world shifted from a model of divine intervention to one of experimentation and investigation (Mennin 2010c). Underlying this was an assumption of order: as knowledge accumulated, it was expected that there would be movement from disorder to order. Phenomena could be reduced into and understood by examining their component parts. Because order was the norm, one would be able to predict an outcome with some precision, and processes could be determined (controlled or predicted) because they would flow along defined and orderly pathways (Geyer et al. 2005). The legacy of this thinking is evident in the way many medical education programs are organized and can even be seen in our approaches to teaching (Mennin 2010c).

The reductionist view, that the whole (or an outcome) can be understood and thus predicted by investigating and understanding the contribution of the constituent parts, is an integral part of the scientific approach that has characterized Medicine for five centuries. The reductionist perspective also dominated educational evaluation throughout a major portion of its short 80-year history as a formal field of practice (Stufflebeam & Shinkfield 2007). This cause-effect approach to analysis requires an assumption of *linearity* in program elements' relationships. That is, changes in certain program elements are expected to have a predictable impact on the outcome. A small change would be expected to have a small impact, a large change a large impact. The assumption of linearity is evident in some popular program evaluation models such as the Logic Model (Frechtling 2007) and the Before, During, and After model (Durning et al. 2007; Durning & Hemmer 2010). Examination of those models shows a logical flow from beginning to end, from input to outcome. The reductionist or linear way of thinking suggests that once the factors contributing to an outcome are known, program success or lack of success in achieving those outcomes can be explained. The cause-and-effect paradigm's impact on several of the evaluation models we describe is clear.

System theory

Although the reductionist approach brought great advances in medicine and even medical education, concern with the approach's limitations can be traced back to at least Aristotle and the dictum that the *"whole is greater than the sum of its parts."* In other words, what we see as a final product—an educational program, a human being, the universe—is more than simply a summation of the individual component parts. The appreciation that an outcome is not explained simply by component parts but that the relationships between and among those parts and their environment (context) are important eventually led to formulation of a *system theory*. In the 20th century, this is often attributed to Bertalanffy, a biologist who proposed a *general system theory* in the 1920s (Bertalanffy 1968, 1972). Although he recognized the roots of his idea in earlier thinking, Bertalanffy's approach focusing on systems was a major step away from the reductionist tradition so dominant in scientific thinking at the time.

Bertalanffy proposed that *"the fundamental character of the living thing is its organization, the customary investigation of the single parts and processes cannot provide a complete explanation of the vital phenomena. This investigation gives us no information about the coordination of parts and processes."* (Bertalanffy 1972). Bertalanffy viewed a system as *"a set of elements standing in interrelation among themselves and with the environment."* (Bertalanffy 1972). Stated another way, the system comprises the parts, the organization of the parts, and the relationships among those parts and the environment; these relationships are not static but dynamic and changing.

In proposing his General System Theory, Bertalanffy noted, *"…there exist models, principles, and laws that apply to generalized systems or their subclasses, irrespective of their particular kind, the nature of their component elements, and the relationships or "forces" between them. It seems legitimate to ask for a theory, not of systems of a more or less special kind, but of universal principles applying to systems in general.… Its subject matter is the formulation and derivation of those principles which are valid for 'systems' in general."* (Bertalanffy 1968) Thus, in his view, an animal, a human being and social interactions are all systems. In the context of this Guide, an educational program is a social system composed of component parts, with interactions and interrelations among the component parts, all existing within, and interacting with, the program's environment. To understand an educational program's system would require an evaluation approach consistent with system theory.

Bertalanffy's proposal (re)presented a way of viewing science, moving away from reductionism, and looking for commonalities across disciplines and systems. Thus, while his ideas about a General System Theory were initially rooted in biology, 20th century work in mathematics, physics, and the social sciences underscored the approach that Bertalanffy proposed: across a variety of disciplines and science, there are common underlying principles.

> A consequence of the existence of general system properties is the appearance of structural similarities or isomorphisms in different fields. There are

> correspondences in the principles that govern the behaviour of entities that are, intrinsically, widely different. To take a simple example, an exponential law of growth applies to certain bacterial cells, to populations of bacteria, of animals or humans, and to the progress of scientific research measured by the number of publications in genetics or science in general. (Bertalanffy 1968)

Finally, General System Theory embraces the idea that *change* is an inherent part of a system. Bertalanffy described systems as either being "closed", in which nothing either enters or leaves the system, or "open", in which exchange occurs among component parts and the environment. He believed that living systems were open systems. Equilibrium in a system means that nothing is changing and, in fact, could represent a system that is dying. In contrast, an open system at *steady-state* is one in which the elements and interrelationships are in balance—still active, perhaps even in opposite or opposing directions, but active nonetheless (Bertalanffy 1968). Furthermore, in an open system, there is equifinality: the final state or outcome can be reached from a variety of starting points and in a variety of ways (much like a student becoming a physician by going through medical school) as contrasted with a closed system in which the outcome might be predetermined by knowing the starting point and the conditions. We believe this view of an open system is consistent with what occurs in an educational program: it is an open system, perhaps sometimes at steady-state, but active.

Since the advent of General System Theory, a number of other theories have arisen to attempt to address the principles across a variety of systems. One such theory, Complexity Theory, is growing in prominence in medical education and thus deserves further consideration of its influence on evaluation choices.

Complexity theory

Linear models based on reductionist theory may satisfactorily explain phenomena that are at equilibrium, a state in which they are not changing. Educational programs, however, are rarely in equilibrium. Medical education programs are affected by many factors both internal and external to the program: program participants' characteristics, influence of stakeholders or regulators, the ever-changing nature of the knowledge on which a discipline is based, professional practice patterns, and the environment in which the educational program functions, to name only a few (Geyer et al. 2005). Medical education programs are therefore best characterized as complex systems, given that they are made up of diverse components with interactions among those components. The overall system cannot be explained by separately examining each of its individual components (Mennin 2010b). In a sense, the program's whole is greater than the sum of its parts—there is more going on in the program (the complex system) than can be explained by studying each component in isolation. This might, in fact, explain the phenomenon in educational research in which much of the variance in the outcome of interest is not explained by factors identified in the system or program: there is more occurring in the program with respect to explaining the outcome than can be fully appreciated with reductionist or linear approaches to inquiry.

Complexity theory and complexity science are attempts to embrace the richness and diversity of systems in which ambiguity and uncertainty are expected. "*Complexity 'science' then is the study of nonlinear dynamical interactions among multiple agents in open systems that are far from equilibrium.*" (Mennin 2010c) *"Complexity concepts and principles are well suited to the emergent, messy, nonlinear uncertainty of living systems nested one within the other where the relationship among things is more than the things themselves."* (Mennin 2010a) Complexity theory allows us to accommodate the uncertainty and ambiguity in educational programs as we think about evaluating them. It actually promotes our understanding of such natural ambiguity as a normal part of the systems typical of medical educational programs. Ambiguity and uncertainty are neither good nor bad but simply expected and anticipated. Evaluating an educational program would therefore include exploring for those uncertainties. In fact, complexity theory invites educators to cease relying on overly simple models to explain or understand complex educational events. *"To think complexly is to adopt a relational, a system(s) view. That is to look at any event or entity in terms, not of itself, but of its relations."* (Doll & Trueit 2010)

The importance of program context is part of complexity theory, helping us to realize the "*work of the environment [in] shaping activity rather than the cognition of practitioners dictating events.*" (Doll & Trueit 2010). In other words, examining a program's success must not only include references to elements related to program participants but also to the relationships of participants with each other and with the environment in which they act and how that environment may affect the participants.

Complexity theory can inform our choice of program evaluation models. For example, the concept of program elements' relationship is prominent in the CIPP evaluation model in which *Context* studies play a critical role in shaping the approach to evaluating program effectiveness and in which program *Process* studies are separate but of equal importance (Stufflebeam & Shinkfield 2007). The need to understand relationships among program elements prompts educators to include a variety of stakeholder views when developing a program evaluation, as each one will reflect key elements of the program components' relationships. The Before, During, After evaluation model (Durning et al. 2007; Durning & Hemmer 2010), described in the literature but not discussed in this Guide, can also be interpreted from the perspective of complexity theory. While there is a linear or orderly nature to that model, it is general and generic enough to allow program planners to envision the rich complexities possible in each program phase and to think broadly about what elements and relations are important within each phase.

Doll suggests that *"... the striving for certainty, a feature of western intellectual thought since the times of Plato and Aristotle, has come to an end. There is no one right answer to a situation, no formula of best practices to follow in every*

situation, no assurance that any particular act or practice will yield the results we desire." (Doll & Trueit 2010) We believe that appropriately chosen evaluation models allow academic managers and educators to structure useful program evaluations that accommodate a program's true complexity. Complexity theory provides a different and useful perspective for choosing an evaluation model that serves program needs more effectively, allowing educators to avoid an overly narrow or simplistic approach to their work.

Common evaluation models

"Educational evaluation" is best understood as a family of approaches to evaluating educational programs. The following discussion of selected evaluation models places them in relationship to the theoretical constructs that informed their development. Thoughtful selection of a specific evaluation model allows educators to structure their planning and to assure that important information is not overlooked.

We will describe four models in this Guide: the familiar experimental/quasi-experimental approach to evaluation; Kirkpatrick's approach; the Logic Model; and the Context/Input/Process/Product (CIPP) model. Educators will find other models in the evaluation literature, but these four are currently in common use and provide clear contrasts among the possibilities offered by models informed by different theories. Each model will be described in some detail, including typical evaluation questions, and what the evaluator might expect when using the model.

The experimental/quasi-experimental models

Experimental and quasi-experimental designs were some of the earliest designs applied as educational evaluation came into common use in the mid-1960s (Stufflebeam & Shinkfield 2007). Arising from the reductionist theoretical foundation, the validity of findings from studies using these designs depends on the evaluator's careful validation of the assumption of linear causal relationships between program elements and desired program outcomes. These designs explicitly isolate individual program elements for study, consistent with the classic reductionist approach to investigation. The familiar experimental and quasi-experimental designs were enormously useful in advancing the biological sciences over the last century (Stufflebeam & Shinkfield 2007). They have proven less useful in the complex environments of educational programs: true experimental, tightly controlled designs are typically very difficult to implement in educational programs as complex as those in medical education. Educators usually need to compare a new way of doing things to the old way of doing things rather than to "doing nothing", so the experimental study's outcomes are usually measures of a marginal increment in value. Quasi-experimental designs are used more often than the true experimental designs that are simply not feasible. Contemporary evaluators shying away from experimental or quasi-experimental designs cite low external validity due to the study design challenges and point to the highly focused nature of such a study's findings.

We now describe and comment on the most commonly used quasi-experimental designs seen in evaluation studies, as those models persist in medical education practice. Educators should be familiar with them in order to make informed choices for their own work.

In the *Intact-Group Design*, learners are randomly assigned to membership in one of two groups. The program being evaluated is used by one of the two groups; the other gets the usual (unchanged) program. The use of randomization is intended to control all factors operating within the groups' members that might otherwise affect program outcomes. Based on the learners' random assignment to groups, the evaluator then acts on the assumption that each group member is an individual replication of the program state (new program or old program). If, for example, each group had 30 members then the analysis would be of $n = 60$ rather than $n = 2$ (groups). For optimal use of this evaluation design, the intact-groups study should be repeated multiple times. If repetition is not feasible, the evaluator/experimenter must continually be alert for unexpected differences that develop between the groups that are not due to the planned program implementation. For example, one group might miss a planned program element due to an event outside the educator's control, such as an unplanned faculty absence. The evaluator/experimenter in this dynamic environment must then attempt adjustments to negate the potential influence of that factor on one group. If the assumption of a linear relationship between the "input" (program type) and the "outcome" is logically defensible and if random assignment to groups has been achieved and maintained, the educator must also guarantee that the programs being compared have been implemented with fidelity and that the impact of unintended events has been equalized.

Evaluators who choose a *Time-Series Experimental Design* study the behavior of a single person or group over time. By observing the learner(s) or group(s) before a new program is implemented, then implementing the program, and finally conducting the same observations after the program, the evaluator can compare the pre- and post-program behaviors as an assessment of the program's effects. This design is similar to the pre/post test design well-known to educators. Time-series studies are most useful when the program is expected to make immediate and long-lasting changes in behavior or knowledge. The number of observations required both pre- and post-program for reliable assessment of changes must be carefully considered. The design does not separate the effects that are actually due to the program being evaluated from effects due to factors external to the program, e.g. learner maturation, learning from concurrent courses or programs, etc. A variation on the time-series design uses different learner groups; for example, learners in early phases of a longitudinal program over several years may be observed to gather pre-program data, while other learners who used the program may be observed to gather post-program data. This requires the evaluator to collect sufficient data to defend the assumption that the "early phase" learners are consistently the same with respect to characteristics relevant to the program even though

the learners observed pre-program are not the same as those observed post-program. For example, all learners in the first year of post-graduate training at Institution X might be observed for two years to collect data about their advanced clinical procedural skills. At the same time, an intensive new program designed to teach advanced clinical procedural skills might be introduced for final-year post-graduate trainees at that institution and data collected after the program for the first two groups (two years) to go through that program. Then the evaluator would compare the early-phase learner data to the post-program learner data, although the groups do not contain the same individuals. The usefulness of this design is limited by the number of design elements that must be logically defended, including assumptions of linear relationships between program elements and desired outcomes, stability of outcome variables observed over a short time period, or (in the case of using different learner groups) sufficient comparability of comparison groups on outcome-related variables.

The *Ex Post Facto* Experiment design, though criticized by some evaluation experts, may be useful in some limited contexts. In this design the evaluator does not use random assignment of learners to groups or conditions. In fact, the evaluator may be faced with a completed program for which some data have been collected but for which no further data collection is feasible. Realizing the weakness of the design, its appropriate use requires analyzing outcome variables after every conceivable covariate has been included in the analysis model (Lieberman et al. 2010). The evaluator must therefore have access to relevant pre-program participant data to use as covariates. When those covariates are even moderately correlated with program outcomes, the program effects may not be detectable with this study design, and a finding of "no effect" may be unavoidable.

What can evaluators expect to gain from experimental and quasi-experimental models? Reductionist approaches are familiar to most medical educators, so experimental and quasi-experimental evaluation studies offer the comfort of familiar designs. The designs do require assumption of linear causal relationships between educational elements and outcomes, although the complexity of educational programs can make it difficult to document the appropriateness of those assumptions. It can also be difficult simply to implement studies of this type in medical education because learning institutions are not constructed like research environments—they rarely support the randomization upon which true experimental designs are predicated. Ethical considerations must be honored when random assignment would keep learners from a potentially useful or improved learning experience. In many educational situations, even quasi-experimental designs are difficult to implement. For example, institutional economics or other realities that cannot be manipulated may make it impossible to conduct an educational activity in two different ways simultaneously. When both feasible and logically appropriate to use these designs, evaluators may choose them when high internal validity of findings is valued over the typically low external validity yielded by experimental and quasi-experimental designs. These designs, used alone, can sometimes provide information about the educational activity's outcomes but cannot provide evidence for why the outcomes were or were not observed.

Kirkpatrick's four-level evaluation model

Kirkpatrick's four-level approach has enjoyed wide-spread popularity as a model for evaluating learner outcomes in training programs (Kirkpatrick 1996). Its major contributions to educational evaluation are the clarity of its focus on program outcomes and its clear description of outcomes beyond simple learner satisfaction. Kirkpatrick recommended gathering data to assess four hierarchical "levels" of program outcomes: (1) learner satisfaction or reaction to the program; (2) measures of learning attributed to the program (e.g. knowledge gained, skills improved, attitudes changed); (3) changes in learner behavior in the context for which they are being trained; and (4) the program's final results in its larger context. To assess learner reactions to the program, evaluators would determine the desired reactions (satisfaction, etc.) and ask the learners what they thought about the program. Learners might be asked, for example, if they felt the program was useful for learning and if individual components were valuable. The second Kirkpatrick "level" requires the evaluator to assess what participants learned during the program. Various designs can be used to attempt to connect the learning to the program and not to other learning opportunities in the environment. Tests of knowledge and skills are often used, preferably with an appropriate control group, to investigate this aspect. A "level three" Kirkpatrick evaluation focuses on learner behavior in the context for which they were trained (e.g. application of knowledge previously gained to a new standardized patient encounter). For example, post-graduate trainees' use of the program's knowledge and skills might be observed in their practice setting and compared to the desired standard to collect evidence for a "level three" evaluation. A "level four" Kirkpatrick evaluation focuses on learner outcomes observed after a suitable period of time in the program's larger context: the program's impact, for example, on patient outcomes, cost savings, improved healthcare team performance, etc.

Kirkpatrick's model has been criticized for what it does not take into account, namely intervening variables that affect learning (e.g. learner motivation, variable entry levels of knowledge and skills), relationships between important program elements and the program's context, the effectiveness of resource use, and other important questions (Holton 1996). The model requires the assumption of causality between the educational program and its outcomes, a reflection of the reductionist linear theories.

What can evaluators gain from using the Kirkpatrick four-level approach? Kirkpatrick's approach defines a useful taxonomy of program outcomes (Holton 1996). By itself, however, the Kirkpatrick model is unlikely to guide educators into a full evaluation of their educational program (Bates 2004) or provide data to illuminate why a program works. Used in conjunction with another model, however, Kirkpatrick's four levels may offer a useful way to define the program outcomes element of other more complete evaluation models (Table 1).

Table 1. Comparison of evaluation models.

CIPP studies	Context studies	Input studies	Process studies	Product studies	
Logic model		Input element→	Activities element→	Output element→	Outcomes element
Kirkpatrick's 4-level model					Learner-related outcomes
Experimental/quasi-experimental models					Linear relationship of intended program outcomes to program elements

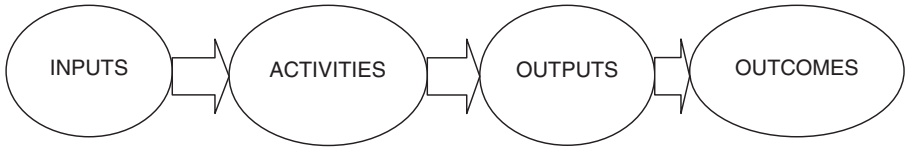

Figure 1. Logic model components.

The logic model

The influence of system theory on the Logic Model approach to evaluation can be seen in its careful attention to the relationships between program components and the components' relationships to the program's context (Frechtling 2007). Though often used during program planning instead of solely as an evaluation approach, the Logic Model structure strongly supports a rational evaluation plan. The Logic Model, similar to the evaluation models already discussed, can be strongly linear in its approach to educational planning and evaluation. In its least complicated form, it may oversimplify the program evaluation process and thus not yield what educators need. With careful attention to building in feedback loops and to the possibility of circular interactions between program elements, however, the Logic Model can offer educators an evaluation structure that incorporates system theory applications into thinking about educational programs. The Logic Model approach to program evaluation is currently promoted or required by some US funding agencies (Frechtling 2007), so it is worth knowing what this approach can offer.

The Logic Model's structure shares characteristics with Stufflebeam's CIPP evaluation model (Table 1) but focuses on the change process and the system within which the educational innovation is embedded. Though its structural simplicity makes it attractive to both novice and experienced educators, this approach is grounded in the assumption that the relationships between the program's educational methods and the desired outcomes are clearly understood. The simplest form of the Logic Model approach may therefore oversimplify the nonlinear complexity of most educational contexts. The Logic Model works best when educators clearly understand their program as a dynamic system and plan to document both intended and unintended outcomes.

The four basic components of the Logic Model are simple to define (Figure 1). The level of complexity introduced into the specification of each component can vary with the evaluator's skill or the program director's resources. When using a Logic Model for program planning, most find it useful to begin with the desired Outcomes and then work backwards through the other components (Frechtling 2007). For complex programs, the Logic Model can be expanded to multiple tiers. Our description will include only the basics of the four essential elements, but details of multi-tiered Logic Models suitable for more complex programs are readily available in texts (Frechtling 2007).

Inputs. A Logic Model's Inputs comprise all relevant resources, both material and intellectual, expected to be or actually available to an educational project or program. Inputs may include funding sources (already on hand or to be acquired), facilities, faculty skills, faculty time, staff time, staff skills, educational technology, and relevant elements of institutional culture (e.g. Departmental or Dean's support). Defining a program's Inputs defines a new program's starting point or the current status of an existing program. Importantly, an inventory of relevant resources allows all stakeholders an opportunity to confirm the commitment of those resources to the program. A comprehensive record of program resources is also useful later for describing the program to others who may wish to emulate it. Readers of this Guide may find it helpful to cross-reference the Input section of the Logic Model to the Input section of Stufflebeam's CIPP model (Table 1). The CIPP model's Input section is a more detailed way of looking at program "inputs" and can be used to expand the construction of the Logic Model's input section.

Activities. The second component of a Logic Model details the Activities, the set of "treatments", strategies, innovations or changes planned for the educational program. Activities are typically expected to occur in the order specified in the Model. That explicit ordering of activities acknowledges that a subsequent activity may be influenced by what happens after or during implementation of a preceding activity. Educators working with complex multi-activity programs are urged to consult a reliable text on the Logic Model for suggestions about developing more elaborated models to meet the needs of their programs (e.g. Frechtling 2007).

Outputs. Outputs, the Logic Model's third component, are defined as indicators that one of the program's activities or parts of an activity is underway or completed and that something (a "product") happened. The Logic Model structure dictates that each Activity must have at least one Output, though a single Output may be linked to more than one Activity. Outputs can vary in "size" or importance and may sometimes be difficult to distinguish from Outcomes, the fourth Logic Model component. In educational programs, Outputs might include the number of learners attending a planned educational event (the activity), the characteristics of faculty recruited to contribute to the program (if, for example, "recruit faculty with appropriate expertise" were a program activity) or the number of educational modules created or tested (if, for example, "create educational modules" were an activity).

Outcomes. Outcomes define the short-term, medium-term, and longer range changes intended as a result of the program's activities. A program's Outcomes may include learners' demonstration of knowledge or skill acquisition (e.g. meeting a performance standard on a relevant knowledge test or demonstrating specified skills), program participants' implementation of new knowledge or skills in practice, or changes in health status of program participants' patients. Outcomes may be specified at the level of individuals, groups or an organization (e.g. changes in a department's infrastructure to support education). Cross-referencing to Stufflebeam's CIPP model's Product section may provide additional ideas for the Outcomes section of a Logic Model (Table 1).

In addition to the four basic Logic Model elements, a complete Logic Model is carefully referenced to the program's Context and its Impacts. Context refers to important elements of the environment in which the program takes place, including social, cultural, and political features. For example, when a governmental or accrediting body mandates a new topic's inclusion in a curriculum, this is a relevant political factor. Learner characteristics may be relevant social factors. Attending to contextual features of a program's environment that may limit or support the program's adoption by others helps planners to identify program elements that should be documented. Impact comprises both intended and unintended changes that occur after a program or intervention. Long-term outcomes with a very wide reach (e.g. improving health outcomes for a specific group) might be better defined as "impacts" than outcomes in a Logic Model approach.

The Logic Model approach can support the design of an effective evaluation if educators are appropriately cautious of its linear relationship assumptions. Typical evaluation questions that might be used in a Logic Model approach include questions like these:

- Was each program activity implemented as planned? If changes from the planned activities were made, what changes were made and why were they necessary?
- Were the anticipated personnel available? Did they participate as anticipated? Did they have the required skills and experience?
- How well did the activities meet the needs of all learners, including learner groups about which the program might be especially concerned?
- What barriers to program implementation were encountered? How was the planned program modified to accommodate them?
- Did faculty participate in associated faculty development? What skills or knowledge did they acquire? How well did they implement what they learned in program activities?
- How did participants in the program activities evaluate the activities for effectiveness, accessibility, etc.?
- What were learners' achievement outcomes?
- How often or how well did learners apply what they learned in their clinical practice?
- How did related patient outcomes change after program implementation?

What should educators expect to gain from using the Logic Model approach? A Logic Model approach can be very useful during the planning phases of a new educational project or innovation or when a program is being revised. Because it requires that educational planners explicitly define the intended links between the program resources (Inputs), program strategies or treatments (Activities), the immediate results of program activities (Outputs), and the desired program accomplishments (Outcomes), using the Logic Model, can assure that the educational program, once implemented, actually focuses on the intended outcomes. It takes into account the elements surrounding the planned change (the program's context), how those elements are related to each other, and how the program's social, cultural, and political context is related to the planned educational program or innovation.

Logic Models have proven especially useful when more than one person is involved in planning, executing, and evaluating a program. When all team members contribute to the program's Logic Model design, the conversations necessary to reach shared understandings of program activities and desired outcomes are more likely to happen. Team members' varied areas of expertise and their different perspectives on the theory of change pertinent to the program's activities and desired outcomes can inform the program's design during this process.

Some potential pitfalls of using the Logic Model should be considered, however. Its inherent linearity (Patton 2011) can focus evaluators on blindly following the Model during program implementation without looking for unanticipated outcomes or flexibly accommodating mid-stream program changes. Evaluators aware of this pitfall will augment the Logic Model approach with additional strategies designed to capture ALL program outcomes and will adapt the program's activities (and the program's Logic Model) as needed during program implementation. A program's initial Logic Model may need to be revised as the program is implemented.

The Logic Model approach works best when the program director or team has a well-developed understanding of how change works in the educational program being evaluated. A program's Logic Model is built on the stakeholders' shared understandings of which strategies are most likely to result in

Table 2. Evaluation questions common to CIPP evaluation studies.

Context	Input	Process	Product
• What is necessary or useful: in other words, what are the educational needs? • What are the impediments to meeting necessary or useful needs? • What pertinent expertise, services, or other assets are available? • What relevant opportunities (e.g. funding opportunities, administrative support) exist?	• What are the potential approaches to meeting the identified educational need? • How feasible is each of the identified approaches, given the specific educational context of the need? • How cost-effective is each identified approach, given the specific educational context of the need?	• How was the program actually implemented, compared to the plan? • How is/was the program implementation documented? • Are/were program activities on schedule? If not, why? • Is/was the program running on budget? If it is/was over or under the planned budget, why? • Is/was the program running efficiently? If not, why? • Can/did participants accept and carry out their roles? • What implementation problems have been/were encountered? • How well are/were the implementation problems addressed? • What do/did participants and observers think about the quality of the process?	• What positive outcomes of the program can be identified? • What negative outcomes of the program can be identified? • Were the intended outcomes of the program realized? • Were there unintended outcomes, either positive or negative? • What are the short-term implications of program outcomes? • What are the longer-term implications of program outcomes? • What impacts of the program are observed? • How effective was the program? • How sustainable is the program? • How sustainable are the intended and positive program outcomes? • How easily can the program elements be adopted by other educators with similar needs?

desired outcomes (changes) and why, so users should draw on research and their own experience as educators to hypothesize how change will work in the program being evaluated. In all cases, however, evaluators should be aware of and explore alternative theories of change that may be operating in the program.

The Logic Model approach will not generate evidence for causal linkages of program activities to outcomes. It will not allow the testing of competing hypotheses for the causes of observed outcomes. If carefully implemented, it can, however, generate ample descriptive data about the program and the subsequent outcomes.

The CIPP (context/input/process/product) model

The CIPP set of approaches to evaluation is described by Daniel Stufflebeam, its creator, as his response to and improvement on the dominant experimental design model of its time (Stufflebeam & Shinkfield 2007). First described in print in 1971, Stufflebeam intended CIPP Model evaluations to focus on program improvement instead of proving something about the program. The usefulness of the CIPP model across a variety of educational and non-educational evaluation settings has been thoroughly documented (Stufflebeam & Shinkfield 2007). Its elements share labels with the Logic Model (Table 1), but the CIPP model is not hampered by the assumption of linear relationships that constrains the Logic Model. An evaluator who understands an educational program in terms of its elements' complex, dynamic and often nonlinear relationships will find the CIPP model a powerful approach to evaluation.

The CIPP approach consists of four complementary sets of evaluation studies that allow evaluators to consider important but easily overlooked program dimensions. Taken together, CIPP components accommodate the ever-changing nature of most educational programs as well as educators' appetite for program-improvement data. By alternately focusing on program Context, Inputs, Process, and Products (CIPP), the CIPP model addresses all phases of an education program: planning, implementation, and a summative or final retrospective assessment if desired. The first three elements of the CIPP model are useful for improvement-focused (formative) evaluation studies, while the Product approach, the fourth element, is very appropriate for summative (final) studies.

Context evaluation study. A CIPP Context evaluation study is typically conducted when a new program is being planned. The associated evaluation questions (Table 2) are also useful when an established program is undergoing planned change or must adapt to changed circumstances. A new leader taking over an existing program, for example, may find thinking through a Context evaluation study helpful. Context studies can also be conducted when decisions about cutting existing programs are necessary. Explicit attention to an educational program's context is essential to effective evaluation and aligns well with complexity theory's emphasis on context.

A CIPP Context evaluation study identifies and defines program goals and priorities by assessing needs, problems, assets, and opportunities relevant to the program. The Context study's findings provide a useful baseline for evaluating later outcomes (Products). When preparing a request for external funding, a program's planning or leadership team can use a good Context study to strengthen the proposal. Because questions about potential impediments and assets are included, a Context evaluation is more inclusive than a conventional "needs assessment", though it does include that essential element.

A number of data collection and analysis methods lend themselves well to a Context study. The evaluator might select from among the following methods, for example, depending on what the situation demands:

- Document review
- Demographic data analysis
- Interviews
- Surveys
- Records analysis (e.g. test results, learner performance data)
- Focus groups

Input evaluation study. A CIPP model Input evaluation study is useful when resource allocation (e.g. staff, budget, time) is part of planning an educational program or writing an educational proposal. An Input evaluation study assesses the feasibility or cost-effectiveness of alternative or competing approaches to the educational need, including various staffing plans and ways to allocate other relevant resources. Incorporating the Input evaluation approach into program development helps to maintain maximum responsiveness to unfolding program needs (context). Building on the associated Context evaluation study, a CIPP model Input evaluation study focuses on how best to bring about the needed changes. A well-conducted Input evaluation study prepares educators to explain clearly why and how a given approach was selected and what alternatives were considered.

A CIPP Input evaluation study formalizes a scholarly approach to program design. When used to plan a new program, an Input evaluation study can also set up clear justification for assigning grant funding or other critical resources to a new program. When applied to a program already in place, an Input evaluation study can help the educator to assess current educational practices against other potential practices. Its focus on feasibility and effectiveness allows a developing program to remain sensitive to the practices most likely to work well.

Identifying and assessing potential approaches to an educational need in an Input study might involve any of the following methods:

- Literature review
- Visiting exemplary programs
- Consulting experts
- Inviting proposals from persons interested in addressing the identified needs

Process evaluation study. A CIPP Process evaluation study is typically used to assess a program's implementation. This type of study also prepares the evaluator to interpret the program's outcomes (see Product study) by focusing attention on the program elements associated with those outcomes. A Process evaluation study can be conducted one or more times as a program runs to provide formative information for guiding in-process revisions. For programs operating in the complex environment typical of medical education programs, this attention to process issues allows an ongoing data flow useful for program management and ongoing effective change. This kind of evaluation study can also be conducted after a program concludes to help the educator understand how the program actually worked. A CIPP Process study explicitly recognizes that an educational model or program adopted from one site can rarely be implemented with fidelity in a new site: contextual differences usually dictate minor to major adaptations to assure effectiveness. The Process evaluation study elicits information about the program as actually implemented. Retrospective Process evaluation studies can also be used to examine often-overlooked but very important program aspects.

The CIPP model's Process evaluation study is invaluable for supporting accountability to program stakeholders. It also allows for the data collection necessary for a program's continual improvement. The "lessons learned" about programmatic processes documented in a Process study are often useful to other educators, even when communication of program outcomes alone may not be all that useful.

An evaluator designing a CIPP Process evaluation study would typically want to use the least-obtrusive methods possible while the program is running. The evaluator might choose from among these methods:

- Observation
- Document review
- Participant interviews

Product evaluation study. The CIPP model's Product evaluation study will seem familiar to most educators because of its focus on program outcomes. What may be more surprising is the breadth of that focus (Table 2). The CIPP Product evaluation study is the one most closely aligned to the traditional "summative" program evaluation found in other models, but it is more expansive. This type of evaluation study aims to identify and assess the program outcomes, including both positive and negative outcomes, intended and unintended outcomes, short-term and long-term outcomes. It also assesses, where relevant, the impact, the effectiveness, the sustainability of the program and/or its outcomes, and the transportability of the program. A CIPP model Product evaluation study also examines the degree to which the targeted educational needs were met. A Product evaluation study may be conducted while a project is running, as interim reports of such a study will be useful for accountability purposes and for considering alternative processes, if warranted by less than desirable findings.

A well-conducted CIPP model Product evaluation study allows the evaluator to examine the program's outcomes across all participants as well as within relevant sub-groups or even for individual participants. Program outcomes (Products) are best interpreted with the findings of the Process evaluation studies in hand: it is possible, for example, that poor implementation (a process issue) might cause poor or unintended outcomes. The art of the Product evaluation study is in designing a systematic search for unanticipated outcomes, positive or negative. To encompass the breadth of a good Product evaluation study, the evaluator might choose from these methods and data sources:

- Stakeholders' judgments of the project or program
- Comparative studies of outcomes with those of similar projects or programs
- Assessment of achievement of program objectives
- Group interviews about the full range of program outcomes
- Case studies of selected participants' experiences
- Surveys
- Participant reports of project effects

What should educators expect if they choose to use the CIPP model? CIPP model studies can be used both formatively (during program's processes) and summatively (retrospectively). Careful attention to the educational context of program is supported, including what comes before, after, or concurrently for learners and others involved in the program, how "mature" the program is (first run versus a program of long standing, etc.), and the program's dependence or independence on other educational elements. The CIPP model incorporates attention to multiple "inputs": learners' characteristics, variability, and preparation for learning; faculty's preparation in terms of content expertise and relevant teaching skills, the number of faculty available at the right time for the program; learning opportunities, including patient census and characteristics and other resources; adequacy of funding to support program needs and leadership support. The CIPP model allows educators to consider the processes involved in the program or to understand why the program's products or outcomes are what they are. It incorporates the necessary focus on program products or outcomes, informed by what was learned in the preceding studies of the program but focuses on improvement rather than proving something about the program. It can provide multiple stakeholders information about the program's improvement areas, interpretation of program outcomes, and continuous information for accountability.

When choosing the CIPP model, educators should be aware that using it effectively requires careful planning. It is most useful if taken up during the planning phases of a new program but may be usefully adopted for retrospective evaluation of a completed program. Multiple data collection methods are usually required to do a good job with CIPP studies, and each data set must be analyzed with methods appropriate to the data and to the evaluation questions being addressed.

Conclusion

Educational programs are inherently about change: changing learners' knowledge, skills, or attitudes; changing educational structures; developing educational leaders; and so forth. The educators who design and implement those programs know better than most just how complex the programs are, and such complexity poses a considerable challenge to effective program evaluation. Academic managers can gain insight into what different evaluation models can do for them by considering the theories that influenced the development of popular evaluation models. The reductionist theory's strict linearity, reflected in the familiar experimental and quasi-experimental evaluation models, may be too limiting to accommodate the known complexity of educational programs. Kirkpatrick's four-level model of learner outcomes also draws on the assumption of linear relationships between program components and outcomes but may be useful in helping evaluators to identify relevant learner outcomes. The Logic Model, often informative during program planning, specifies the intended relationships between its evaluation components and may require constant updating as a program evolves. The Logic Model's grounding in systems theory prompts adopters to incorporate the program's context in evaluation studies, making it more inclusive than earlier evaluation models. Stufflebeam's CIPP model is consistent with system theory and, to some degree, with complexity theory: it is flexible enough to incorporate the studies that support ongoing program improvement as well as summative studies of a completed program's outcomes. Medical educators can choose from these individual models or a combination of them (Table 1) to develop an evaluation model adequate for their programs.

Declaration of interest: The authors report no declarations of interest. The views in this Guide are those of the authors and do not necessarily represent the views of the United States Government, the United States Air Force, or other federal agencies.

Notes on contributors

ANN W. FRYE, Ph.D., is Director of the Office of Educational Development and Assistant Dean for Educational Development at the University of Texas Medical Branch, Galveston, Texas. She specializes in educational evaluation and research in the medical education context.

PAUL A. HEMMER, M.D., MPH, is Professor and Vice Chairman for Educational Programs, Department of Medicine, Uniformed Services University of the Health Sciences, F. Edward Hebert School of Medicine, Bethesda, MD, USA

References

ACGME. 2010a. Accreditation council for graduate medical education: glossary of terms. Accreditation Council for Graduate Medical Education. Available from: http://www.acgme.org/acWebsite/about/ab_ACGMEglossary.pdf

ACGME. 2010b. Program director guide to the common program requirements. Accreditation Council for Graduate Medical Education. [Accessed 31 January 2011] Available from: http://www.acgme.org/acWebsite/navPages/nav_commonpr.asp

Bates R. 2004. A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. Evaluat Program Plan 27:341–347.

Bertalanffy L. 1968. General system theory: Foundations, development, application. New York: George Brazilier, Inc.

Bertalanffy L. 1972. The history and status of general systems theory. Acad Manage J 15:407–426.

Cook DA. 2010. Twelve tips for evaluating educational programs. Med Teach 32:296–301.

Doll Jr WE, Trueit D. 2010. Complexity and the health care professions. J Eval Clin Pract 16:841–848.

Durning SJ, Hemmer P, Pangaro LN. 2007. The structure of program evaluation: An approach for evaluating a course, clerkship, or components of a residency or fellowship training program. Teach Learn Med 19:308–318.

Durning SJ, Hemmer PA. 2010. Program evaluation. In: Ende J, editor. ACP teaching internal medicine. Philadelphia: American College of Physicians.

Frechtling J. 2007. Logic modeling methods in program evaluation. San Francisco: John Wiley & Sons.

Geyer R, Mackintosh A, Lehmann K. 2005. What is complexity theory? Integrating UK and European social policy: The complexity of Europeanisation. Abington: Radcliffe Publishing.

Goldie J. 2006. AMEE education guide no. 29: Evaluating educational programmes. Med Teach 28:210–224.

Hawkins RE, Holmboe ES. 2008. Constructing an evaluation system for an educational program. In: Hawkins RE, Holmboe ES, editors. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby, Inc.

Holton E. 1996. The flawed four-level evaluation model. Hum Res Dev Quart 7:5–21.

Kirkpatrick D. 1996. Revisiting Kirkpatrick's four-level model. Train Dev 1:54–59.

LCME. 2010. Functions and structure of a medical school. Standards for accreditation of medical education programs leading to the M.D. degree. Washington, DC: Liaison Committee on Medical Education. [Accessed 31 January 2010] Available from: http://www.lcme.org/standard.htm

Lieberman S, Ainsworth M, Asimakis G, Thomas L, Cain L, Mancuso M, Rabek J, Zhang N, Frye A. 2010. Effects of comprehensive educational reforms on academic success in a diverse student body. Med Educ 44:1232–1240.

Mennin S. 2010a. Complexity and health professions education. J Eval Clin Pract 16:835–837.

Mennin S. 2010b. Complexity and health professions education: A basic glossary. J Eval Clin Pract 16:838–840.

Mennin S. 2010c. Teaching, learning, complexity and health professions education. J Int Assoc Med Sci Educat 20:162–165.

Musick DW. 2006. A conceptual model for program evaluation in graduate medical education. Acad Med 81:759–765.

Patton M. 2011. Developmental evaluation: applying complexity concepts to enhance innovation and use. New York: Guilford Press.

Stufflebeam D, Shinkfield A. 2007. Evaluation theory, models, & applications. San Francisco: Jossey Bass/John Wiley & Sons, Inc.

Woodward CA. 2002. Program evaluation. In: Norman GR, Van Der Vleuten CP, Newble DI, editors. International handbook of research in medical education. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Recommended readings

In addition to those listed in the References, the following resources are recommended for learning more about theory and other evaluation models:

Abma TA. 2005. Responsive evaluation: Its meaning and special contribution to health promotion. Evaluat Program Plan 28:279–289.

Guba E, Lincoln YS. 1989. Fourth generation evaluation. Beverly Hills: Sage Publications.

Donaldson SI. 2007. Program theory-driven evaluation science: Strategies and applications. Mahwah, NJ: Erlbaum.

Shadish WR, Cook TD, Leviton LC. 1991. Foundations of program evaluation: Theories of practice. Newbury Park, CA: Sage.