Planning Evaluation Through the Program Life Cycle

Mary Ann Scheirer¹ and Thomas Schwandt ²

Introduction

Linking evaluation methods to the several phases of a program’s life cycle can provide evaluation planners and funders with guidance about what types of evaluation are most appropriate over the trajectory of social and educational programs and other interventions. If methods are matched to the needs of program phases, evaluation can and should generate a continuously evolving body of evidence that is useful to program managers and administrators as they develop, improve, deliver, and spread more effective programs.

Contributors to this Forum discuss a framework for evaluation planning that outlines the uses of evaluation methods during an extended period of program development, testing, delivery, and dissemination. This framework may also help organize the many diverse strands of theory and practice in the evaluation field, by suggesting that different methods emphasized by diverse writers about evaluation should be matched to the different stages of program development and delivery. An initial article provides an introductory overview of the framework. Five commentaries follow; each discusses the framework from a particular perspective—an academic perspective, a view “from the trenches” of human service delivery, perspectives from current and prior government evaluators, and a retrospective on several evaluations funded by a major health-focused foundation. These commentaries provide examples of the application of the framework, and, in some instances, point out limitations in this approach to evaluation.

¹ Scheirer Consulting
² University of Illinois

Corresponding Author:
Mary Ann Scheirer, Scheirer Consulting
Email: MaryAnn@scheirerConsulting.com
Expanding Evaluative Thinking: Evaluation Through the Program Life Cycle

Mary Ann Scheirer

Introduction

Recent evaluation policy statements, including American Evaluation Association (AEA)’s “An Evaluation Roadmap for a More Effective Government” (American Evaluation Association [AEA], 2010) have suggested that evaluative thinking and planning ought to be part of an expanded life cycle for social and educational programs. By employing a variety of methods that are appropriate for the different stages or phases of a program, evaluation can become an integrated managerial function in which data are continuously collected and used for decision making and program improvement. A program’s “life cycle” includes: (a) program planning and development; (b) sometimes, testing the program for causal effectiveness; (c) normal continuous program delivery; and (d) dissemination and replication of effective programs to other organizations. Along the same lines, program planners and policy makers have recently advocated evidence-based practice (EBP) among health care providers and prevention specialists. For EBP, researchers first develop interventions that are rigorously tested for efficacy, which they release for dissemination into widespread use. At the local level, program managers using those EBP interventions should employ continuous monitoring to assure effective implementation and responsiveness to community needs.

Little guidance is currently available to either practicing evaluators or policy makers about what it might mean to promote evaluation throughout a program’s life cycle. Traditionally, most evaluation guidance calls for conducting discrete studies that produce one or more reports, often conducted by a person external to the organization that is implementing the target program. Moreover, there is as yet no unifying framework that brings together the various methods for evaluation that have been advocated by diverse evaluation approaches—for example, participatory and empowerment evaluation (Fetterman, 1996; Fetterman, Kaftarian, & Wandersman, 1996; Fetterman & Wandersman, 2005; Whitmore, 1998), theory-based evaluation (Chen, 1990; Donaldson, 2007; Donaldson & Gooler, 2002), performance measurement (Hatry, 1999; McDavid & Hawthorn, 2006; Wholey, 1997) and many others—to suggest how they might be particularly appropriate at various stages of a program’s life cycle. Although the developmental stage of a target program might be a consideration in planning specific evaluation studies, there has not been a systematic focus on planning for evaluation as a continuous function throughout program development, impact testing, and later routine delivery stages, with multiple types of data collected to inform each phase.

This framework grows out of my career-long work on evaluating program processes other than their causal effectiveness, beginning with studies to evaluate implementation processes and quality (Scheirer, 1981, 1987; Scheirer & Roberts-Gray, 1988), then expanding into a fuller description of various process evaluation methods (Scheirer, 1994), and continuing with my more recent work on evaluating program sustainability (Scheirer, 2005; Scheirer & Dearing, 2011). While planning a set of evaluation workshops for the U.S. Agency for International Development (USAID) in South Africa in 1997, an administrator asked, “When should we evaluate our programs?” I replied, “All the time! Evaluation methods are useful during every stage of program development,

1 Scheirer Consulting
implementation and delivery.” So we developed a brief workshop amplifying this theme for the local USAID staff members. Later, I presented an overview of this topic in an invited plenary panel at the Annual Meeting of the Eastern Evaluation Research Society in 2008 (Scheirer, 2008).

The focus for this life cycle framework is on a program as the evaluand, with a program defined as the development and delivery of a specified set of activities intended to address one or more indentified outcomes, especially in the human services arena (Newcomer, Hatry, & Wholey, 1994). A program may also be termed an “intervention” especially when it is intended to intervene in some aspect of the lives of the recipients, to address their needs and enhance opportunities for productive lives and healthy behaviors. In contrast to my intended usage here, sometimes the term “program” is simply a funding stream, such as a line item designating a program in the federal budget that allocates money for a broad purpose. Those programs may encompass many different types of activities in different locations, rather than a specified set of activities or components. Using the term “program” for such a diverse set of local activities makes federal-level evaluation more difficult and likely to yield ambiguous outcomes, for there is not a well-developed and well-defined evaluand to be evaluated. In addition, when broad scale systemic change is intended, the intervention strategy might include multiple “programs”—sets of activities—focused on diverse forces interacting in an environment (Leischow & Milstein, 2006; McLeroy, Bibeau, Steckler, & Glanz, 1988; Williams & Imam, 2007).

Evaluation is broadly defined here as the systematic collection and use of data—both qualitative and quantitative—to aid users in developing knowledge about and managing a targeted set of activities. Evaluation may also blend into types of studies often termed research, when considered from a life cycle perspective that includes needs assessment, problem definition, and identification of risk factors. An assumption underlying this definition is that evaluation is conducted for program improvement and, ultimately, for social betterment, rather than focusing primarily on the generation of knowledge per se, which is often seen as the purpose of research. It is also important to acknowledge that not all parts of this evaluation life cycle are necessarily conducted within the same organization. Evaluators or researchers who perform the initial development and efficacy testing may be in quite different places and roles from evaluators who assess ongoing program delivery, as pointed out by several commentators in this Forum.

Previous Approaches That Have Influenced This Framework

Many streams of theory and practice in evaluation that are too numerous to specify individually have influenced this framework. It is similar to Glasgow’s RE-AIM approach for health-related interventions, which proposes research on five aspects of a program’s potential development and spread: its Reach, its Effectiveness, its Adoption by new users, its Implementation, and its Maintenance (Glasgow, Vogt, & Boles, 1999). Glasgow’s framework seems to focus primarily on designing multiple types of data collection within a single research initiative, rather than suggesting diverse types of evaluation at different points in a program life cycle. Similarly, EBP in the health field assumes a process over time of using basic research to develop interventions, which are then tested for efficacy before being disseminated for widespread delivery by other implementers (Health Affairs, 2005). This approach is complemented by practice-based research in which the realities and constraints of health care practice are communicated to the developmental researchers and included in dissemination strategies for interventions, to increase the two-way linkage between research and practice (Olsen, Aisner, & McGinnis, 2007).

Focusing on evaluation in relation to program life cycles is not entirely new to the evaluation field, but perhaps has been overlooked amidst recent debates about impact assessment. In 1980, Cronbach and his colleagues published a critique of the strong evaluative emphasis on testing the impacts of programs in relation to predetermined goals. In contrast, they called for recognizing four phases of program maturity, the developmental or “breadboard” phase, a “superrealization” phase
to assess what the program can accomplish “at its best,” a third stage of a substantial field trial, then fourth, the established program stage (Cronbach et al., 1980, pp. 108–109). Ten years later, Chelimsky analyzed ways that social science and evaluation can contribute to governmental decision making, and suggested, “The immaturity of a field is as much a problem for determining the effectiveness of an ongoing program as it is for policy or program development” when there is little theoretical or empirical foundation, measuring tools have not been developed, and the interventions are not systematically described (Chelimsky, 1991, p. 227). More recent work by Trochim and his colleagues at Cornell has cited the “medical research to translation” paradigms in health care to emphasize that effective medical interventions are based on a lengthy period of diverse types of trials before evidence-based interventions are achieved. Building on this conceptual model, Trochim recommends a phased approach in matching the fit of evaluation methods to program development-and-delivery stages throughout program life cycles (Cornell Office for Research on Evaluation, 2009; Trochim, 2009). Other evaluators have suggested using evaluation methods across the program life cycle, but to my knowledge have not developed this approach into a full framework for evaluation and program planning.

This framework has also benefited from the wisdom and experience of Eleanor Chelimsky, who has described three major purposes for evaluation:

A. Evaluation for accountability to funders and other stakeholders to ensure that funding is being used appropriately; performance measurement is often done for this purpose. Accountability can also include assessing efficiency, such as costs per unit of service.

B. Evaluation for causal knowledge about a program or intervention to generate strong evidence that the intervention independently causes the intended outcomes; also called impact or efficacy evaluation.

C. Evaluation for program improvement by using data for managing a program; being utilization-focused; often involves tracking and using data about shorter-term outcomes, rather than longer-term efficacy (Chelimsky, 1997).

The following discussion of evaluation during various program life cycle stages will amplify how these three purposes receive differential emphasis at different time points by evaluators in different roles. There is not a one-to-one correspondence between these purposes and life cycles stages, but it is important to clearly specify the intended purposes and audiences for evaluation at each stage.

Outline of the Evaluation Life Cycle Framework

This conceptual framework for evaluation assumes a developmental perspective over time for program creation, testing, and delivery. To be carried out effectively, the development and testing work may require 2–5 years, especially depending on the length of time for one cycle of program delivery, and on how many iterations of program development are needed before intended outcomes are produced. Although the framework is presented as occurring in phases, in reality these phases are not necessarily discreet, nor always easily identifiable. Often a program may be funded for widespread use, even without the program development or efficacy testing stages described here. When evaluation is then attempted on these inadequately developed programs, the results are often disappointing, with findings showing neither clear outcomes nor documenting anticipated effectiveness. Thus, this framework is both a conceptual model that is abstracted from reality and a suggested normative process for working toward more effective human service programs, with accumulation of evidence across organizations and over time.
Program Development Phase

Ideally, new or modified programs will be developed systematically, rather than funded on an ad hoc basis with an initial focus on broadscale service delivery. Programs that are effective in producing desired outcomes for beneficiaries often start with an extended period of problem analysis and intervention development. Even new program components within existing programs may benefit from an evidence-based developmental process. To contribute to developing the innovative programs, evaluative work during this phase would first focus on needs assessment and understanding of the problem being addressed in collaboration with content area specialists. With their backgrounds in research methods and understanding evidence, evaluators can bring evidence from prior literature into discussions with content specialists who are developing a new intervention. Methods suggested by evaluation for needs assessment could include: (a) qualitative data collection (e.g., focus groups, open-ended interviews) from potential clients to provide in-depth understanding of their needs and strengths, as well as showing the contexts of their lives; (b) analysis of existing databases, such as prior surveys or population databases for risk factors associated with the problem, as well as data on the scope of the problem area (how many and what types of people are affected?); and (c) literature review and synthesis about what is known in the problem area, what interventions have been tried in the past and with what results? For what aspects of the problem area are new interventions needed or is more evidence needed about the effectiveness of previously developed interventions?

As the program developers create new program activities for an innovative intervention, evaluation methods can again be helpful by using formative feedback about pilot delivery to determine whether the new intervention is acceptable to potential clients and other stakeholders, if it is compatible with the cultures and expectations of those intended to participate, and if recipients begin to make the intended changes. Examining cultural compatibility includes assessing whether the intended new intervention fits with the normal ways of doing things among both intended program delivery staff and participants expected to make behavioral change. Development of an intervention logic model at this stage will make clear the intended activities and outcomes, and link them to the prior literature and underlying assumptions of the new program (Frechtling, 2007; Kellogg (WK) Foundation, 2004; McLaughlin & Jordan, 1999; Wholey, 1983). If stakeholders have diverse views about the content of the logic model, or if some parts of it are ambiguous, that may signal that further work is needed to develop and clarify the intervention.

As the new program undergoes initial implementation, usually in a pilot site, evaluability assessment might be used to undertake a fuller exploration of stakeholders’ assumptions and aspirations for the program, and whether the program activities are logically related to one another and to expected outcomes (Wholey, 1987). Are program activities sufficient in scope and intensity to bring about expected outcomes, when the program starts collecting preliminary evidence about its intended outcomes, especially its short-term outcomes? The Robert Wood Johnson Foundation recently employed this approach by using evaluability assessments, expert panels, and synthesis to identify innovations that have the potential for effectiveness in combating children’s obesity (Leviton, Khan, & Dawkins, 2010). If the intended intervention/interventions involve systemic change in a broader system, then including a fuller agenda of process evaluation would be helpful to describe the actual intervention/interventions utilized, the players involved in the initiative, the extent of implementation, and the initial changes or other outcomes in other parts of that system.

This program development stage may require 2–5 years of work to create a promising intervention that is ready for the next, impact assessment phase. Evaluation data collection during this stage should be ongoing to bring prior evidence, to bear, and to inform learning about diverse aspects of the developing intervention. This stage focuses most heavily on Chelimsky’s third purpose of evaluation for program improvement, as data inform both intensive needs assessment and provide initial evidence about how the pilot intervention works in practice.
An example of using multimethod evaluation for development of a complex program in a challenging environment is illustrated by the development of the MANAS program in India (Chatterjee et al., 2008). The researchers identified strong local needs for treatment of common mental health problems, such as depression and anxiety disorders, yet were faced with a scarcity of trained personnel and resources to treat them. Extensive development work over 15 months included: (1) reviewing relevant literature from countries facing similar resource limitations; (2) developing an intervention using lay health counselors trained to deliver stepped care for patients identified by a screening tool; (3) a consultation phase to interview primary care doctors and potential intervention recipients about the likely acceptability and feasibility of the intended intervention; and (4) two rounds of trial delivery using lay health counselors to identify likely barriers to program implementation, with data collected via both quantitative process indicators and interviews with staff and recipients. These trial delivery rounds identified implementation problems such as lack of private space in local clinics for the mental health counselors to meet with patients and low rates of follow-up by patients for subsequent appointments. After each trial round, modifications in program details and components (such as adding yoga sessions to increase the cultural acceptability) were made to overcome the barriers identified. The final multicomponent program was then assessed in a cluster randomized trial, which found lay mental health workers to be more effective in reducing the mental disorders of intervention patients after 6 months, than was the usual care provided for these conditions, especially in public facilities (Patel et al., 2010). This program is now engaged in a widespread dissemination effort across other areas of India.

Testing the Program for Causal Efficacy

With a fully developed and prior evidence-based intervention available, including organizational experience with promoting successful implementation, the program is ready for more intensive testing for causal efficacy, the traditional and historic focus of much evaluation methodology. Testing for causal efficacy refers to a comparative assessment of the measured outcome/outcomes versus an estimate of what those outcomes would have been in the absence of that program. The objective of the evaluative work shifts to providing causal evidence that the outcomes that occurred during the development phase are indeed caused by the intervention/interventions, rather than possibly due to other biasing factors, such as historic change or client maturation. This phase is likely to require one or more well-designed evaluative research studies in order to set up and carry out an experimental or quasi-experimental design, with methods that are well described in the evaluation methodology literature (Boruch, 1997; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). This phase clearly encompasses Chelimsky’s second purpose of evaluation for causal knowledge.

Reichardt recently provided an extensive framework of methods for estimating program effects by pointing out that strong estimates can be accomplished by randomizing settings, times, or outcome variables, as well as by randomly assigning individual recipients to experimental and control groups (Reichardt, 2011). Although a randomized controlled trial (RCT) is often considered one of the strongest potential designs for unbiased estimates of program or treatment effects, the circumstances under which that design can be effectively employed are relatively stringent (cf. AEA, 2008; Boruch, 1997; Trochim, 2009). These criteria include:

- The program is well defined and has an articulated program model, such as one developed during the developmental phase described above.
- High quality (i.e., valid and reliable) outcome measures are available and are feasible to collect.
- Prior evaluation has shown that the program is capable of producing desired change in those outcomes.
- There is ambiguity whether outcomes occurring are caused by the program; that is, a clear need is present for the stronger evidence from a randomized trial.
The program has been implemented consistently and with high fidelity during the developmental phase; measures will be used during the impact assessment phase to measure fidelity of implementation.

- The scope of the study has sufficient statistical power to detect the anticipated extent of change.
- The random assignment can be implemented and maintained.
- Ethical and human subject protections have been approved and are in place.
- The study has both the financial resources and research expertise to be carried out with high quality.

In sum, this extensive set of criteria for a valid RCT is not likely to be present within small nonprofit agencies that are often the recipients of human service program grants. Collaboration between knowledgeable program managers and expert researchers is likely to be needed in order for impact assessment to employ a randomized design.

Other research designs can also provide strong evidence of causal impact, even without an RCT. In some instances, it may be possible to use data from a naturally occurring experiment, for example, when different states are trying different intervention approaches for welfare reform and have similar existing data sources for assessing outcomes over time. Or a longitudinal interrupted time-series design might provide strong evidence of causal effectiveness, if the needed long-term and frequently collected data are available (Orwin, 1997). Evidence is also accumulating from place-randomized designs that use larger-scale units, such as schools, housing complexes, hospital intensive-care units, or entire villages, especially to assess the effectiveness of larger scale or systemic change initiatives that involve multiple changes within those units (cf. Boruch, 2005).

Of course, any impact evaluation must also include strong fidelity of implementation evidence in order to insure a fully delivered and received intervention, or conversely, to analyze impact findings in relation to variations in implementation. Experience with “black box” impact evaluations in the early years of the profession found that outcome results showing no differences between treatment and control groups may be uninterpretable if systematic data were not collected about the extent and fidelity of implementation. Was the program not implemented well (implementation failure), or was it not more effective than its comparison, even when well implemented (theory or program development failure)?

A comparative causal evaluation may be more difficult to design for large-scale systemic change initiatives that may include different program interventions for different types of participants or places (Victoria, Black, Boerma, & Byce, 2011). But causal evidence may be less salient if the intended outcomes are occurring, and the pattern of outcomes documented by the outcome data matches the hypotheses derived from theoretical underpinnings of the systemic change (pattern-matching evaluation; Marquart, 1990; Trochim, 1985). The underlying premise of many environmental or systemic change interventions is that a number of components interact over time to produce the desired change, rather than a linear process of intervention producing change in individual recipients, which is the more common program model for individually focused human service programs. The specific interactions of components in each time period and place may be unique to that situation, so that a generalizable causal model is not feasible.

After several or numerous impact evaluations of the same or similar interventions are available, evaluation/research synthesis becomes possible and desirable, using meta-analysis, if feasible, or more traditional literature review methods (Labin, 2008). Meta-evaluation for assessing the utility of diverse research methodologies may also be desirable. The synthesis and comparison of findings from multiple evaluations of the same or similar interventions produce a stronger base of evidence than any one study alone, or this synthesis might raise questions about the generalizability of results from a specific evaluation. Such bases of evidence are now accumulating in the intervention reviews produced by many federal agencies, as well as by the Cochrane Collaboration for health studies and...
the Campbell Collaboration for several human service areas. As evidence is accumulated about an intervention or set of interventions addressing a problem area, it is also vital to collect and analyze cost information about program delivery. What does it cost to use those interventions in widespread practice?

Continuous Delivery Phase

Whether or not programmatic interventions being used in normal program delivery have been based on this type of extensive development and impact assessment cycle, it is very desirable to include evaluative data collection as feedback within normal program management processes. This type of evaluation has several names that are frequent used, including data for monitoring program delivery; data used for continuous quality improvement; and local performance measures with feedback to staff members. Of Chelimsky’s three major purposes for evaluation, the key purpose here is evaluation for program improvement, rather than to document causal impact. The evaluative data may also be needed for accountability, to provide evidence to funders and other stakeholders that intended outcomes are occurring (Chelimsky, 1997). However, the program activities might also be modified further, as staff members learn what steps help increase participation and attendance, which activities are most engaging to intended beneficiaries, and so forth. If the modifications to an evidence-based program are extensive, especially if they modify the core program components of interactions with clients, it is essential to also collect outcome data to ensure that the modified program does produce the longer-term outcomes evidenced by the causal efficacy study.

In many cases, this internal program evaluation is the responsibility of local program managers and staff, rather than conducted by a professional evaluator (cf. Sonnichsen, 2000). The most useful type of lower-cost evaluation for local program delivery is likely to be process data to examine implementation delivery, extent of participation, and short-term outcomes of the intended program. With data about local variations in program delivery, staff can make incremental changes in a continuous quality improvement cycle that helps generate the desired longer-term client benefits. This advice contrasts with an initial evaluation emphasis on longer-term outcomes, which are often more expensive and more difficult for local staff to collect, and less under the control of program staff to achieve.

Ideally, the local evaluation will be linked conceptually to related evidence-based programs from prior impact evaluation research and/or to the research-based standards for practice in the content area. When strong evidence demonstrates that defined sets of program activities do lead to specific desired outcomes, then it is less necessary for each local program replication to focus on measuring longer-term outcomes. By analogy, after an immunization is developed and shown to prevent a specific disease, it is not necessary for each local health clinic to track longer-term outcomes of the incidence of that disease. But each clinic ought to use its monitoring data to ensure that all its clients receive that immunization, at the recommended dosage and time points. When the local program is based on a program model developed previously, then it is essential to monitor program delivery for fidelity of implementation, as well as to assess near-term outcomes of the program for the particular population group being served. Is the program working here? Why or why not? In addition, it is important to evaluate whether programs that receive short-term external funding are sustained after that initial funding ends (cf. Scheirer, 2005; Scheirer & Dearing, 2011).

This type of evaluation is very likely to use utilization-based evaluation (Patton, 1997) and participatory methods, so that program deliverers and other stakeholders learn to use methods needed to collect and use systematic data as feedback for managing their programs. For example, the “Getting to Outcomes Framework” of Wandersman and his colleagues focuses on training program managers to use data to address 10 questions needed for program self evaluation (Chinman, Imm, & Wandersman, 2004). When an evaluation uses rapid feedback methods for quality improvement, local program managers and staff are likely to become more engaged in using the data for program
improvement. For example, one independent consultant reports using “real time evaluation memos” every several months when working with a client, to provide fresh data at times when critical decisions are made (Hwalek & Williams, 2011). However, it is usually not feasible for local program managers and staff to undertake the intensive evaluation needed for the Phase 2 impact assessment, unless they are linked with the methodological expertise needed and the additional financial support required for impact evaluation.

Collecting and using systematic data for performance management is also a type of evaluation used during continuous program delivery (Hatry, 1999; Hendricks, 2010). Performance management goes several steps further than earlier emphases of performance measurement by focusing on the use of data for real-time quality improvement. Initial emphasis of the performance measurement movement was on providing high-level data about outcomes to stakeholders, especially for accountability. But broad-level indicators were found not very useful for local program management and improvement when they were not built on a base of program delivery indicators. A full set of performance measures is needed that includes data about the details of local program delivery, in order for it to be useful for program quality and outcome improvement, via performance management.

“Both program evaluation and performance measurement are part of the performance management cycle . . . Both are intended to be part of the feedback loop that reports, assesses, and attributes outcomes of policies and programs” (McDavid & Hawthorn, 2006).

Multiple types of data collection might be especially needed if the intended project involves change in a large-scale system of interacting organizations and actors. No one type of data is likely to adequately capture the multiple processes involved in systemic change (Eoyang & Berkas, 1999). Ideally, there will be publically available data series to track the longer-term outcomes intended for systemic change efforts, such as re-incarceration rates for prisoner release initiatives or ongoing behavioral risk factor surveys for large-scale substance abuse prevention campaigns among youth (such as the U.S. Youth Behavioral Risk Factor Surveys, sponsored annually by the U.S. Centers for Disease Control & Prevention). In that case, local program evaluation could focus on whether the intended activity components of the initiative are carried out with the intended scope of activity, along with measurement of initial behavioral changes that are expected to lead to the broader, perhaps unmeasured systemic changes.

**Dissemination and Replication**

While local program managers are implementing and monitoring their delivery of program activities, other actors are often engaged in disseminating programs that have been developed via the program development and impact assessment phases described above. A focus on the “spread” of effective programs brings with it the necessity of evaluating diffusion and marketing methods, and collecting data about dissemination results for local-level program adoption, implementation fidelity, and longer-term sustainability of evidence-based programs developed via the prior phases. Extensive current work sponsored by the U.S. National Institutes of Health on the “translation” of research into broader scale practice is investigating the processes involved in successfully disseminating and implementing evidence-based practices into widespread use (cf. NIH Conference, 2011). All the methods cited above might be usefully employed to assess dissemination and replication, with a focus now on activities effective in fostering the spread of evidence-based programs, rather than the development and testing of the program activities, per se. Several theoretical models and bodies of evidence are relevant to the focus on dissemination, replication, and sustainability, including diffusion of innovations theory developed by Rogers and his colleagues over many years (Greenhalgh, Robert, Macfarlane, Bate, & Kyriakidou, 2004; Rogers, 1995), social marketing approaches (Andreasen, 1995), and more recent sustainability theory (Johnson, Hays, Center, & Daley, 2004; Pluye, Potvin, & Denis, 2004; Scheirer & Dearing, 2011).
The health field has also begun to sponsor practice-based research to bring practitioners’ concerns and feasibility issues to the attention of researchers, which might stimulate a new round of intervention development and testing. The generalizability of evidence from a controlled impact assessment trial can then be assessed by summarizing evidence from the practice-based research in multiple locations and with multiple types of clients.

Some Implications of a Life Cycle Perspective for Evaluation

This life cycle framework has focused specifically on evaluating programs with well-developed sets of activities and identified intended outcomes, across four life cycle phases. Figure 1 provides a summary of the framework, showing the major types of evaluation applicable to each phase. In several places, the discussion was extended to consider broader initiatives that include many different types of activities intended to stimulate systemic change.

This framework also calls for evaluative planning that goes beyond the frequently used formative–summative distinction (originally formulated by Scriven, 1991). Although formative evaluation methods are certainly needed during the program development phase, methods often viewed as formative—such as process evaluation of program implementation and short-term outcomes—are also integral to evaluation during later phases, especially ongoing program delivery. Further, careful measurement of program implementation is essential for inclusion in causal efficacy evaluation to avoid uninterpretable “black-box” results. A variety of evaluative methods are needed at all program life cycle phases that are not well described by the formative–summative dichotomy.

The framework might also apply to evaluating policy change, especially by calling attention to the need for formative evaluation methods to help develop policy change initiatives before they are enacted by a legislative body. However, if a policy change is enacted without an extended period of developmental work and evaluation, it may be difficult to carry out a systematic study of its intended impact, at least within that jurisdiction, if the policy covers all individuals within that jurisdiction and appropriate comparisons are not available. Therefore, national policy changes may be more difficult to evaluate for their causal impacts than corresponding changes at the state or local levels, for
which ongoing “business as usual” in other similar states might serve as comparisons. (However, an interrupted time-series design might be employed in this situation, if the necessary long-term data series for a relevant outcome is available.) The key point for evaluators working with evaluands other than programs is to consider carefully the appropriateness of specific evaluation methods in relation to the characteristics of the entity being evaluated.

This approach also means that the funders of both programs and evaluation need to plan for the fit of evaluation methods with the life cycle stage of the intended program. For example, over many years, the Edna McConnell Clark Foundation has focused its efforts on developing, assessing, and then disseminating programs to help young people from low-income backgrounds to become productive adults (Edna McConnell Clark Foundation, 2011). It may waste funders’ resources to create large initiatives with many grantees addressing a problem area, without first funding the developmental work using intensive formative evaluative methods to design, then to test potential program activities for effectiveness. Similarly, funders that require causal impact evaluation or long-term outcome effectiveness reporting from local grantees, without providing the necessary resources and expertise to collect that data, are likely to be disappointed with the quality of evaluation produced and to frustrate grantees with requirements that are beyond their capabilities.

A further implication is that evaluators’ roles and responsibilities differ substantially at different program life cycle stages. During the program development stage, evaluators need to be flexible and nimble in quickly designing and collecting data that will be immediately useful to the next design stage of activity. This is likely to apply both to internal evaluators who are employees of the organization in which they work, and to external consulting evaluators who are helping to provide data to illuminate the design work. This stage might not even have a comprehensive evaluation plan developed in advance, as data needs may occur continually in relation to program events or immediate outcomes. In contrast, the impact assessment phase is most often carried out by external evaluators with strong research and statistics skills, along with research support staff experienced in the many detailed activities needed to work with both program managers and recipients of the intervention being tested. Or the impact assessment might be initiated by a university-based research group who possess the necessary evaluation expertise, and then collaborate with local partners for program delivery of the intervention being tested. For this phase, a carefully developed evaluation design is essential, with detailed planning for maintaining its integrity, and for sustaining the intended activities of both program staff and recipients.

At the continuous local delivery phase, evaluators’ roles again shift to the local agency, either working continuously to produce the needed monitoring data, or often, training and coaching local program staff to collect and analyze their own feedback data. An evaluation design may be very useful to formalize the intended data collection and responsibilities, but the design might also be readily changed to accommodate new data needs. For evaluation of dissemination and replication, most current data-oriented work is being done by research organizations or university-based evaluators, perhaps because data collection is required from many local agencies or other actors. The implications of these diverse evaluation roles are substantial when designing training for evaluators: different career trajectories may require quite different skill sets.

The appropriateness of differing types of evaluation at different time points at a program life cycle also depends heavily on the resources available for evaluation—of both financial support and the expertise needed to design and carry out different types of evaluation.¹ If few resources are available in times when public funds are stretched, the most useful data are likely to be monitoring-types of evaluation and/or ongoing local performance measures. These can provide quick feedback about whether the intended activities are being carried out, and whether intended recipients are participating with sufficient consistency to bring about the anticipated short-term outcomes. Larger-scale, more resource-intensive evaluation research studies for impact assessment may need to be deferred until times when the necessary funding is feasible. This might also be a good opportunity for public—
private partnerships, with foundations or other private sources supplying funding for causal impact studies of interventions that have preliminary evidence for promising outcomes, and public agencies or their grantees collaborating by working directly with the intended clients in carrying out the intended interventions.

Conclusion

Many different types of evaluation and research are desirable to illuminate the multiple processes involved in improving the effectiveness of service delivery throughout a program’s life cycle. No one type of evaluation alone will do the job. No single organization is likely to engage in all these types of evaluation. In order for the multiple pieces of this puzzle to come together to improve the whole picture of successful service delivery, it is vital to accumulate findings across the multiple phases of program development, impact assessment, and normal service delivery, as well as for the dissemination of evidence-based programs into widespread use. Systematic evaluation is vital for each of these phases, with evaluation incorporated as an ongoing function of program development and delivery, rather than viewed primarily as discrete studies of causal impact. Evaluators may also serve as linking agents by bringing program funders and managers into touch with relevant evidence about their intended work, both from prior studies and new data collections.

This vision for more useful evaluation means that funders of program development and its associated evaluation may need a longer time horizon than the usual 3-year grant cycle, to allow adequate developmental work before evaluating causal impacts. It also means that all those involved in service delivery—front line program managers and staff, funders, and evaluators—should frequently consult the diverse types of evidence about the social sectors in which they work. Evaluators and local program managers should not have to work alone; their work should be informed by and stand on the shoulders of both the giants and the everyday practitioners of growing bodies of interventions to improve the lives of intended clients.

Note

1. The author is indebted to an EVAKTALK reader who pointed out that the types of evaluation most feasible at particular time points may differ between public and private agencies.

References


The notion that program phase, or alternatively program stage or program maturity, can and should guide evaluation choices, is an attractive one. It seems intuitively compelling that the chief way/ways that evaluation can best contribute will vary across the life cycle of a program. For example, when a program is being planned or is a relative newborn, evaluators probably can best contribute with actions such as needs assessment, formative evaluation, the explication or co-construction of a program theory, and literature-based assessment of the plausibility of key links in the program’s theory of change. In contrast, when the program is more mature, summative forms of evaluation are probably more appropriate, especially if there are questions about program continuation, cessation, or expansion (Chen, 2005; Scheirer, 2011).

The preceding paragraph describes a relative simple stage model, whereby formative evaluation is prescribed for a less mature program while summative evaluation is prescribed for a more mature program. This simple stage model, in one variant or another, is probably familiar to most experienced evaluators. More elaborate versions also exist, though they may be less familiar. In one notable example, Chen (2005) discusses different kinds of evaluation activities he believes are appropriate for four different stages of a program. He also discusses how two additional considerations should affect the choice of evaluation activities: Whether the evaluation is intended to meet

\[1\] Pennsylvania State University
information needs inside and outside the program, and a set of possible trade-offs (between comprehensiveness, rigor, cost, and timeliness).

In short, then, the general idea of attending to program maturity when making evaluation choices is not new, as Scheirer (2011) acknowledges. Nevertheless, Scheirer’s life cycle model has notable merits. It provides a concise, accessible, and thoughtful integration of a wide array of evaluation options. This integration is organized around a sensible linkage of evaluation methods and approaches to different program phases. Relative to most if not all previous life scale models, Scheirer’s model offers a more comprehensive view of program life cycle, focusing on four phases: program development, testing for causal efficacy, continuous delivery, and dissemination and replication. In contrast, even the relatively detailed, Chen (2005) stage model focuses on (a more elaborated version) of the first steps of Scheirer’s model.

Despite its notable benefits, there are also some limitations of the model, and of life cycle models more generally. Alternative models also exist that do not focus on program life cycle but do link the choice of evaluation options to other factors. These may serve as complements or alternatives to life cycle models. In addition, Scheirer’s model helps point the way to future directions that might further enhance the value of life cycle models.

Strengths and Limits of, and Alternatives to, Life Cycle Models of Evaluation

Appropriate application of a life cycle model, as highlighted by Scheirer’s version, should help bring about several important benefits, including the timely application of evaluation aimed at program improvement, thereby increasing the likelihood the programs will contribute to desired outcomes; the avoidance of premature summative evaluation, thus reducing the risk of ending potentially effective program models that are still developing; better matching of evaluation activities to the situation; avoidance of one-size-fits-all thinking about evaluation; and helping funders have realistic expectations and move toward more sensible, longer-term funding both of program development and of evaluation.

At the same time, limitations apply to a life cycle model. Perhaps, the primary general concern involves the extent to which the underlying life cycle model applies well, both to the program being evaluated and to the information needs of important stakeholders. Take a phenomenon familiar to many evaluators: Sometimes the wheels fall off a program that previously was moving along nicely. For example, a program may have strong leadership, enthusiastic staff, adherence to a sensible program plan, and correspondingly positive effects in its second year; however, by Year 4 or 5, management and the best staff members may have left for better or more secure positions, remaining staff members are burnt out, and overall adherence to the original program plan has been lost to low motivation and to inadequate training of replacement staff. To be sure, a life cycle model can handle such setbacks. The typical linear presentation of such models, however, can lead readers to assume that program age is a good indicator of program stage, and similarly to expect that the typical program progresses relatively smoothly from one life cycle phase to the next.

Another way that practical realities may intrude on life cycle-based planning occurs when the relevant decision space precludes timeframes of the length that Scheirer describes. For example, in the run up to reauthorization of welfare reform, states received waivers that allowed testing of alternative approaches to welfare. Image that a state engaged in Phase 1, program development-oriented evaluation for so long that summative evaluation findings were not available prior to Congressional debate about reauthorization. If so, adherence to a life cycle model could have precluded the legislative process from being informed about the merit and worth of that state’s new approach to welfare.

In addition, one can speculate that different versions of program life cycle may be required for different circumstances. Federally funded multisite pilot programs might have different phases, or different ways in which a given phase unfolds, and thus different evaluation needs, relative to
programs initiated at the state level. This may especially be true regarding dissemination and replication. Likewise, programs (or policies) that begin large in scale may in general have different cycles than those that start small. For programs with universal coverage, for instance, one might imagine that evaluation aimed at testing efficacy should be both preceded and followed by evaluation related to continuous delivery.

In short, a life cycle model may not fit evaluation well when a stage model is not a good fit to program history and to key information needs. One direction for future work would be to assess the conditions under which various variants of a life cycle model best fit. Scheirer’s Figure 1 is a helpful generic life cycle model, but more specific versions might work better in certain conditions. Another future direction would include discussion, and perhaps empirical research on the extent to which life cycle models are more or less appropriate, relative to alternative kinds of models, in guiding contingent evaluation planning. Models that emphasize program stage or life cycle are not the only kind of evaluation theories that emphasize contingent decision making (Mark, Greene, & Shaw, 2006). One alternative focuses on stakeholders’ planned use of evaluation, with Patton’s (2008) utilization-focused evaluation a leading example. Another alternative focuses on a more policy analytic assessment of likely contribution of various evaluation activities to social betterment. Mark, Henry, and Julnes (2001) fall into this category. They discuss the use in evaluation planning of assessing whether there is, for example, an upcoming decision about program adoption (or replacement, or cessation), as in the case of welfare reauthorization, versus a relatively stable period of program continuation. Again, a fruitful future direction might involve discussion, and perhaps empirical research about the relative benefits and fit of these different kinds of models under different circumstances. Probably, the best approach for contingent selection from among evaluation options will involve choosing whichever kind of model fits current conditions best, as well as integrating across these different types of contingent models. A slogan might be “Even the contingencies are contingent.”

Referring back to the Scheirer model, a comment Scheirer made might also help in identifying when the model best applies: “This framework is both a conceptual model that is abstracted from reality, and a suggested normative process for working toward more effective human service programs.” Using Alkin’s (2004) language, Scheirer indicates that hers is meant to be both a descriptive model (capturing what evaluators actually do) and a prescriptive model (describing what evaluators, and in this case funders, should do). When funders have followed, or are open to following, a program life cycle close to that summarized in Figure 1, then the model presumably is a good guide to evaluation planning. In contrast, if funding and decision making deviate greatly from the assumptions of the life cycle model, then presumably it is less helpful as a guide to contingent decision making about evaluation options. One notable message of the Scheirer article is that evaluators should encourage funders to follow a more reasonable model for program development, testing, and dissemination. Of course, practical considerations (e.g., changes in the party in power) and potential lack of political will likely make this a long-term effort, especially in the public sector.

There is much more to be admired and considerably more to be said about Scheirer’s article than space allows here. One important point, deserving of more attention than can be given here, is how issues of external validity or the generalizability of findings play out in the life cycle model. Phase 2 tests of causal efficacy are not likely to answer all current and future questions related to generalizability, including the generalizability-related questions that arise in taking a pilot program to scale. To what kind of circumstances can the program be disseminated with equal or greater effectiveness? What aspects of the program are central to its success and need to be implemented with fidelity at new sites? What forms of adaptation facilitate cultural responsiveness and successful implementation in new settings (or times), and which adaptations deviate from key ingredients of the program model and inhibit success? Which evaluation procedures will better address these and related questions? Perhaps future elaborations, expanding on Scheirer’s excellent contribution and drawing on additional experience with life cycle evaluation, will be able to address such questions in more detail.
Note

1. To be fair, it should be noted that these other types of contingency theories also have their limits. In the present article, however, only the life cycle or program stage approach is assessed in terms of its strengths and weaknesses.

Commentary on “Expanding Evaluative Thinking: Evaluation Through the Program Life Cycle”

Ariana Brooks

In this commentary, the phases of Scheirer’s program life cycle are discussed from the perspective of an internal evaluator for a large American nonprofit organization, emphasizing the extent to which the phases match the reality of most nonprofit contexts and budgetary constraints. The fundamental issue that a nonprofit has to confront is that when funding becomes available, the evaluation process is often guided by the government entity. When looking at designing the evaluation process, the nonprofit is often restricted to utilizing limited resources to develop a data system that will meet requirements set by the funding source.

As Scheirer states the ideal first step in initiating a program includes a needs assessment and literature review. Both research and practice often illustrate that this is the most ideal approach. Unfortunately, most nonprofits approach program development with one thought in mind; namely, what the funder is going to fund. In most instances, it is not the organization deciding the need; rather it is the government funder. However, if the organization or funder is developing a new program design, then a needs assessment and literature review is reasonable and ideal.

1 HeartShare Human Services of New York
Another important issue that arises in the program development phase is the capacity, in terms of human resources and expertise, of an organization to conduct evaluation activities before the initiation of a program. Often, for nonprofits, the resources are not available until after the program has begun. For example, some funding is contingent upon services billed. Thus, the program must already be operational to receive funding. Additionally, the expertise needed to conduct literature reviews, for example, may also be limited. Although some organizations look to universities to provide evaluation expertise, most nonprofits, especially small community-based organizations (i.e., those with an annual budget of less than five million), often utilize program managers who have other responsibilities for the task of developing their evaluation processes. The evaluation then often consists of data from case record reviews and/or consumer surveys, rather than planned comparisons or randomized controlled field trials.

Additionally, there may be major needs to develop evaluation tools related to specific nonprofit services and populations. For instance, funders could provide support to help an agency develop a valid and reliable tool to assess satisfaction for nonverbal individuals. While there may be available measures or research on working with nonverbal individuals, testing measures, methods, or questions that specifically relate to the program’s services and client population could be useful in developing the most practical tool for assessing those services.

Another significant matter that is not considered in the development phase of the Scheirer framework is discussion of an organization’s commitment to evaluation as evidenced in its culture. Specifically, how can an organization develop a culture of accountability that would include and strengthen the value of the evaluation process when developing a program? One can institute evaluation processes throughout the program life cycle, but unless there is a culture that reinforces the importance of the evaluation within the organization, it will not succeed and probably do more harm than good. For example, it could lead to a misinterpretation of the overall program value or staff dissension.

In the second phase of the perspective, testing for causal efficacy, Scheirer emphasizes that utilizing experimental or comparative methods is not practical for many nonprofits, which matches our experienced reality. In my organization, an experimental test would have to be a pilot project, rather than “normal” program delivery. Moreover, any effort to conduct a randomized design or quasi-experimental design would be highly unlikely unless the government funder provides funds in the budget to conduct this type of causal analysis. For many programs, the government will focus its attention on designing a series of performance benchmarks that an organization would be required to meet in order to receive continued funding. For example, recently New York City recalled their original Requests for Proposals (RFP’s) for all of their child welfare programs to address their process for reviewing RFPs. Specifically, there were several issues regarding the validity of the benchmarks used to score proposals. But, there was no attempt to address causal efficacy beyond setting performance benchmarks that programs must meet to maintain their contracts.

The other research designs discussed by Scheirer (e.g., longitudinal interrupted time series design) are viable methods for establishing causal effectiveness; however, these might have a much broader scope and not be relevant to many nonprofit programs with a much more defined scope and population. In my experience, utilizing a general causal model in human service programs is not feasible. When dealing with individuals who bring unique characteristics and circumstance to the table, distinguishing causal relationships for each individual becomes difficult especially when operating with limited resources. Consequently, many programs rely on funders to dictate the scope of the program rather than conducting their own research. So we suggest that Scheirer’s proposed approach should be adopted by government funders who have the funding capacity to investigate possible causal relationships underlying observed outcomes or performance measures.

The third phase of continuous delivery is where there is the most consistency with what nonprofits currently experience. Again, using the recent New York City child welfare example, the RFP
was designed to spell out the standards and performance benchmarks that would have to be met. Further, the city agency clearly identified specified data systems that would provide ongoing information to assess how well the programs were meeting their benchmarks, as well as to identify best practices to help programs that were not meeting performance measurement targets.

Finally, when looking at dissemination and replication, I believe that few nonprofits have the time or the resources to replicate or evaluate their programs in other sites. As illustrated by Scheirer, more emphasis should be placed with government agencies and other funders to work on dissemination and replication and to provide resources for programs to do so. While there are some venues for programs to share positive results of their evaluation work, there is a need for a more systematic and analytical approach. For example, the New York City child welfare department utilizes an accountability workgroup, consisting of internal evaluators from the organizations they fund, to assess evaluation activities and results at a system-wide level. This group regularly meets to review results from various performance measures and to provide feedback on the validity of the tools and measures used to assess program performance and outcomes.

Scheirer’s perspective is logical and ideal. If funders and nonprofits did evaluation throughout a program’s life cycle as the framework suggests and used evidence about intervention efficacy from other sources, then both program staff and program funders would gain meaningful information such as the strengths and challenges of the services offered to specific client populations. This would ensure program quality and that the needs of the client population are met. Overall, linking evaluation methods to the different phases of a program’s life cycle can provide evaluation planners and funders with guidance on what types of evaluations are most appropriate over the trajectory of programmatic interventions, as Scheirer states. If funders begin to shift their priorities in the direction of Scheier’s perspective, nonprofits have much to gain.

A Lifer’s Perspective on Life Cycle Evaluations

George F. Grob1

I have had the good fortune to be an evaluation lifer. My evaluation career began as an operations research analyst at the U.S. Department of Defense in 1969 but matured in the U.S. Department of Health, Education, and Welfare (now Health and Human Services). There, I served for 15 years as a program and policy analyst within the Office of the Assistant Secretary for Planning and Evaluation, another 17 in the Inspector General’s Office of Evaluation and Inspections, and one and a half years with the Citizens’ Healthcare Working Group. After I retired from Federal service, I have continued to evaluate public programs as an independent consultant. My remarks here are based on my experience and observations as a Federal Government evaluator over the years.

Life cycle considerations in social program evaluation are now common enough to warrant the excellent and careful synthesis that Scheirer has crafted in this Forum. I cannot add much to what

1 Center for Public Program Evaluation
she has done. However, I am up to suggesting some possible new paths forward. For that, it is useful to begin by pointing out that few evaluators stay with a program for its life. We analyze only life segments—program births, their shakedown years, their maturation, their lingering old age, seldom their death. Therefore, I thought it might be useful to suggest possible approaches to improve evaluation of programs at various points in their lives.

Programs at Birth

A tectonic shift occurred in evaluation practice with the introduction of evaluability assessments and evaluation program logic models. Few practicing evaluators today start an evaluation without an examination of a program’s stated purpose; its inputs, outputs, near and intermediate outcomes, and impacts; its theory of change; it is program activities (the black box of program gears); moderating external factors; extant performance indicators; relevant body of evaluation; and current program status.

However, there is still a lot missing from how program managers and evaluators plan for and assess new social programs. In this regard, we could learn a trick or two from the Defense sector. As noted above, that is where I began my career, and remarkably, my evaluative work there was all about life cycle evaluation.

In the 1960’s, the Department of Defense was struggling with cost overruns on major weapon systems—including jet fighters, frigates, fleet-based missile defense systems, and nuclear submarines and their nuclear missiles. Cost overruns still occur, but in those days, they were whoppers. To its credit, the Department of Defense went after them systematically, changing the way it controlled development and production, tightening cost and schedule controls, and making plans based on life cycle considerations and cost estimates. The procedures they created then are still in use today, although probably better. In retrospect, the unglamorous innovators and leaders of these management and oversight systems were true pioneers in thinking about a program’s life cycle.

They systematized at least two ways of evaluating and directing, and thereby improving, the successful launch and implementation of defense programs: (1) identifying and scheduling of developmental program stages for evaluation, analysis, and intervention and (2) life cycle cost estimating during these planning stages.

Program development stages

A typical defense weapon system now goes through well defined and managed stages such as conceptual development, systems design, engineering design, prototype testing, and initial production. At each stage, data are assembled and analyzed and presented to a high level board for a go/no-go decision.

No comparable stages and associated evaluations and decision mechanisms are available for social programs. The closest we come is the legislative authorization process (see below). But this lacks the discipline of defense weapons systems development. Newly launched social programs may be periodically evaluated but not against an expected life cycle. Perhaps, evaluators should turn their creative minds to developing frameworks for evaluating new social programs not just at launch, but at well-defined developmental decision points.

Life cycle costing

The Defense community uses life cycle cost estimates not just at the beginning of the development of a new defense system but also at each of the stages mentioned above and then periodically thereafter. Of course, the initial estimates do not always hold up. But the persistence of such estimating informs policy making and is a critical aspect of the evaluation cycle of these systems.
Weighing the benefits of programs against their costs seems logical enough. Indeed, some evaluators are now incorporating cost analysis into their practice. However, life cycle cost estimating is not widely used in the development of social programs. This might well be worth pursuing more vigorously.

**Programs’ First Steps**

The Government Accountability Office (GAO) and many Inspectors General (IG) perform early implementation reviews. Typically, evaluators wait 6 months to a couple of years after a program starts. Before that evaluation results would likely be that things are chaotic “out there,” and many things are delayed. So after a suitable delay, the evaluators go out to the field to collect data. Of course, they read status reports and budget summaries. But the real story is out in the field.

These evaluators gather data through a combination of interviews, written surveys, and preliminary (but frustratingly incomplete and unreliable) program data. They use formal schedule targets, enrollment data, budget and expenditure analysis, and available performance standards. All this enables them to compare how things are to the way they are supposed to be. At the same time, semi-structured interviews with grantees, beneficiaries, and field-level program officials provide deeper insights. Typically, evaluators ask basic questions such as: What is working? What isn’t? What are your biggest challenges? How are you dealing with them?

The evaluators likely find some grantees who are making good progress and others who are not. Finding out what the top performers are doing to overcome obstacles can be very beneficial to those who are still having trouble. Some grantees may have started sooner than others. Sharing their stories of problems and solutions will be very beneficial for the newest grantees. The evaluation results are useful not only for field level grantees and officials but also for headquarters officials as well. The early years are when they can actually make adjustments to program rules and expectations.

A surprising benefit of these early implementation reviews is unanticipated insights into how programs work. Many examples show how improvements were discovered during such studies: That effective local sting operations can identify convenience stores selling tobacco to minors, that unmarried fathers visiting their children at the hospital immediately after birth were willing to acknowledge paternity, that alcohol content of wine coolers sold near schools was illegible because of lack of color contrast and unconventional locations on labels, and that first responders were unable to communicate with one another during broad emergencies because of incompatible communications equipment. While these important factors were well known at some local levels, they were not initially known by national policy makers.

There is no need for GAO and IG’s to monopolize these early implementation reviews. Perhaps, more evaluators can offer this service to their clients.

**Programs Growing Up**

Program performance indicators are widely used today. Their importance is also recognized in the Government Performance and Results Act. However, there has been less recognition of the fact that many programs grow in effectiveness. The performance indicators therefore need to grow with them. The life cycle analogy is particularly germane here. One expects far more of an 18-year-old than of a 4-year-old, and in all aspects of growth, including physical fitness, performance in schools, socialization, and self-sufficiency. The indicators society uses to assess performance change over the child’s years.
The same is, or should be, true for government programs. What we expect from current child support enforcement program, for example, is far more advanced than when the program started. Modern methods of paternity establishment, requirements for health insurance coverage of the children, mandatory employer withholding of support payments, and standards for support enforcement orders were hardly dreamed about when the program first started. The performance indicators and evaluation criteria for such programs have grown and evolved as well.

Another good example is the Federal Runaway and Homeless Youth program. In the early days, its performance indicators focused on the number of youth served and the physical conditions of its homes, centers, and shelters. Gradually, the program adopted measures of services provided and has moved to longer-term outcomes, measuring the percent of children with “safe landings” (care in families, rather than returns to the street when they leave the program).

As evaluators, we seldom talk about program maturation, instead focusing on how well programs get started and when they might be ready for an impact evaluation. That way of thinking might get in our way of life cycle evaluation of programs that gradually mature, almost unnoticeably passing from one phase of life to another. This deserves more thought.

**Rejuvenation**

Most government programs do not die. They mature as described in the previous section, undergo transformations, or get swallowed up into other programs as a result of the legislative reauthorization process. At the Federal level, except for entitlement programs like Medicaid, Medicare, Social Security, and unemployment insurance, most programs are periodically considered for reauthorization (typically every 3–5 years). Life cycle evaluations could be particularly useful if they were available at reauthorization time, when decisions are made about the next cycle of a program’s life. However, the calendar works against linking fresh evaluations to reauthorizations. Policy making for the next reauthorization cycle begins about 18 months in advance, and a thorough evaluation might need to begin 2 years before that, greatly reducing program experience to evaluate. One practical solution is to keep the evaluation and research pipelines full and then synthesize what is known at reauthorization time.

Recently, a new way has been emerging to solve this problem. It is reflected in some recent legislation such as that of the President’s Emergency Relief Plan for AIDS Reduction program. In that case, the Congress requested an evaluation in the law that authorized the program. In essence that law said, “Look, we’ll be back to look at this again in five years. This is what we will want to know then.” The law ensures that a policy relevant evaluation gets done. This process is starting to catch on. We see more and more examples of the Congress requiring evaluations for decision making later on. Evaluators who have an “in” with legislators may be able to suggest this approach.

**Ramifications**

There are three bottom lines to the analysis above: (1) While the field of evaluation has in recent years developed highly innovative and useful tools for examining a program at is “birth” and after it has matured, equally creative evaluative concepts could be valuable during program startup, maturation, and rejuvenation stages. (2) Policy makers, program managers, and evaluators could all benefit from the creation of a developmental model for programs that clearly identifies key decision points and the evaluative information that is brought to the table when such decisions are made. (3) Life cycle cost estimating could be valuable at all stages of a program’s life, not just at the beginning.
Evaluation Purpose and Use: The Core of the CDC Program Evaluation Framework

Thomas J. Chapel

More than 20 years after the publication of *Utilization Focused Evaluation* (Patton, 1978) and its strong emphasis on use of findings, one would expect the “self-evident truths” of our discipline to include that evaluations across a program’s life cycle should be connected and that evaluation purposes and uses will vary along that life cycle. And, yet, like Scheirer, I find they are not. And I appreciate her proposed framework as a way to reinforce those key points.

The article refers to several approaches and frameworks. I would like to offer another one. The underlying philosophy of Centers for Disease Control (CDC)’s *Framework for Program Evaluation in Public Health* (Centers for Disease Control and Prevention, 1999) includes, at least implicitly, all the key points and emphases of her life cycle framework. Her paper is timely because in a current project reflecting on the first decade of use of the CDC Framework, I am finding that some of my cherished assumptions about what the Framework means to say are either not shared by other users and/or not being communicated as they should be. Some of the points in Scheirer’s proposed life cycle framework will help us make these core points more clearly.

Briefly, the CDC Framework (see Figure 1) was developed in order to motivate more program evaluation, but, as importantly, to ensure that the program evaluations that were completed made a difference. Hence, the last step, “use findings and share lessons learned”. And central to getting use is our third step, “focusing the evaluation and its design” in which evaluators identify purpose and intended user of the evaluation and its findings. The resulting focus informs all steps that follow: data collection, analysis, and reporting. Moreover, the focus discussion is fed by the two preceding Framework steps; broad engagement (not merely identification) of stakeholders, and development of a clear and consensus program description.

In case the sequence of steps does not drive home the point that evaluations must be matched to the situation in order to get use of findings, the middle of the graphic includes a box referencing the four clusters of evaluation standards (Centers for Disease Control and Prevention, 1999). We intended to convey that utility, feasibility, propriety, and accuracy frame and influence decisions made at each of the six steps. Utility and feasibility are seen as especially important to setting the right focus: Utility to ensure the focus aligned with the purpose and the needs of the user, feasibility to ensure that stage of development, among other factors, was taken into account as a reality check (Centers for Disease Control and Prevention, 1999; U.S. Department of Health and Human Services [USDHHS], 2005). Like Scheirer’s proposed life cycle framework, we emphasize that the purpose of the evaluation will vary by stage/phase of the program; that variations in purpose will influence the choice of evaluation questions, data collection methods, and analyses approaches; and that there is some evaluation worth conducting (i.e., some plausible evaluation focus) throughout the life of the program.

However, our recent project reflecting on use of the CDC Framework after its first decade (MacDonald & Chapel, 2009) found that this notion of integrated steps and standards often gets missed or lost. Users complete the early steps in sequence but fail to see their interrelationships. Their evaluation focus, for example, may be plausible but appear disconnected from the stakeholder engagement and program description that preceded it. Likewise, the standards are sometimes seen as

---

1 U.S. Centers for Disease Control and Prevention
30 specific rules rather than as four guiding principles or “orientations.” To counteract these tendencies, in recent trainings and presentations, I have been emphasizing that the early steps are iterative; that stakeholder and focus are not only early steps in a sequence but undergird decisions at all other steps; and, especially, that the standards are general principles to be attended to during all six steps.

But other colleagues have placed the fault elsewhere, and their solution lands very close to the points in Scheirer’s proposed life cycle framework. My colleagues, while acknowledging the validity of my points about the steps and standards, assert that these points are made too late in the evaluation cycle. They agree that purpose and use vary with stage but argue that discerning purpose and use must occur before one begins even the first step of the Framework. With this idea in mind, some colleagues (Dunet, 2011), constructed a matrix of eight potential purposes (see Table 1). They then showed how each purpose might engender different choices at each framework step. Their purposes are easily subsumed under Chelimsky’s (1997), three core purposes—accountability, causal knowledge, program improvement. And, they crosswalk nicely to the four phases in Scheirer’s life cycle framework. Table 2 elaborates the figure in Scheirer’s article by expanding the purposes at each phase and distinguishing the purposes from the data collection approaches.

Besides helping me better affirm the points that are central to the CDC Evaluation Framework, the proposed life cycle framework helps with the following potential challenges and misinterpretations that plague our discipline:

- Reinforces a commitment to continuous quality improvement. Although the approach emphasizes the program life cycle, the same points work equally well once we enter the phase of
execution—seeing planning, performance measurement, and evaluation as complementary and integrated processes.

- Accurately positions formative evaluation. In some recent EvalTalk exchanges on formative versus summative evaluation, I was chided, perhaps rightly, for confining formative evaluation to the early stages of a program life cycle. Others pointed out that there are formative and summative questions at every stage of a program’s life. If we see evaluation proceeding across the life cycle then we see room for formation and summation all along the way, although the mix and emphasis will shift by phase.

- Accurately positions evaluability assessment. My favorite sentence in this article is “When should we evaluate our programs? I replied. ‘Always!’” Our CDC Framework emphasizes exactly that. I have not liked the term “evaluability assessment” for fear that some will conclude, after the assessment, that there is nothing to evaluate or the program is “not ready” to evaluate. The life cycle framework makes clear that there are always some productive evaluation questions to ask.

- Clarifies the ongoing debate about causal attribution and rigor of methods. Clearly at some stages of the life cycle, demonstrating cause is the central question, but it is not the only question nor

## Table 1. Potential Purposes of a Program Evaluation

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability</td>
<td>Compile data to demonstrate to stakeholders that your program is functioning as expected</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Routinely examine data to track expenditures, accomplishments, and other key indicators to guide program management</td>
</tr>
<tr>
<td>Improvement</td>
<td>Identify operational strengths and weaknesses to devise ways to improve a program</td>
</tr>
<tr>
<td>Understanding</td>
<td>Identify essential program elements and opportunities for streamlining or enhancing the program</td>
</tr>
<tr>
<td>Replicability</td>
<td>Determine to what extent your program can be well implemented in different settings</td>
</tr>
<tr>
<td>Judgment</td>
<td>Assess whether the program provides a worthwhile return on investments of time, money, and other resources</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Assess how effectively the program is achieving its desired outcomes</td>
</tr>
<tr>
<td>Development</td>
<td>Exploring, building, and testing new ways to meet a public health need</td>
</tr>
</tbody>
</table>

## Table 2. Elaboration of the Program Life cycle Framework

<table>
<thead>
<tr>
<th>Stage</th>
<th>Program Development</th>
<th>Testing for Causal Efficacy</th>
<th>Continuous Delivery</th>
<th>Dissemination and Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purposea</td>
<td>Development</td>
<td>Understanding (Causal) Knowledge</td>
<td>Accountability Monitoring Improvement Understanding</td>
<td>Accountability Monitoring Replicability Judgment</td>
</tr>
<tr>
<td>Methods/Approaches</td>
<td>Formative research Qualitative methods Literature reviews/ systematic reviews Needs assessments</td>
<td>RCT Quasi-experimental design Program theory evaluations Meta-analyses and syntheses Systemic change evaluations</td>
<td>Performance measurement Process evaluation Outcome monitoring</td>
<td>Performance measurement Process evaluation Outcome monitoring</td>
</tr>
</tbody>
</table>

* Purposes are based on Figure 1 in the article on which this commentary is based; refined based on Chelimsky (1997) and Dunet (2011).
even the primary one at other stages in the life cycle. And indeed, if we do not address the other questions in due course, the causal question becomes hard to answer or meaningless.

Finally, I offer the following modest enhancements or clarifications to the life cycle framework, especially viewing this from the perspective of someone engaging in public health evaluation in very applied settings: (1) The first two phases may operate in different sequences. Sometimes program development is followed by tests for causal efficacy. Just as often, we may review the literature for efficacious interventions, then develop the program, and then implement, paying particular attention in delivery to implementation fidelity. (2) The cycle is depicted as linear, but I know Scheirer means to convey that it is recursive at least in the first 3 steps, where experience of the initial group of projects perhaps sends us back to the program development and testing phases, or replication identifies the situations where the program works or does not work, again sending us back to program development and efficacy testing.

Reading this article was both cathartic and constructive. Cathartic in that someone put into words the challenges I have been experiencing in program evaluation. Constructive in that the proposed life cycle framework provides a lens through which to better understand the cause of those challenges and to refine my approach in a way that surmounts them.

References


MacDonald, G., & Chapel, T. (2009, June). *Program evaluation meets the real world: 10 Years of the CDC evaluation framework*. Paper presented at AEA/CDC Summer Evaluation Institute, Atlanta, GA.


Lessons on the Program/Evaluation Life Cycle

Mary Geisz¹, Molly McKaughan¹, and Laura Leviton¹

From the perspective of the Robert Wood Johnson Foundation (RWJF), the nation’s largest philanthropy devoted exclusively to health and health care, coordinating evaluation with each stage in the program life cycle, as Scheirer suggests, is a key to effective evaluation. RWJF, with an evaluation staff of 10 professionals plus support staff, has designed and funded evaluations of most of its programs since its earliest days, in the 1970s. While most evaluations are performed via a contract or grant from RWJF to an independent outside evaluator, a few are conducted by the outside office managing the program.

Evaluations have been used to measure the strategies and outcomes of funded programs; to examine the processes and implementation of ongoing programs in order to improve them; and also to judge the sustainability and possibility of replication of an array of programs. As the RWJF Anthology: To Improve Health and Health Care states, RWJF has learned some lessons about the program/evaluation life cycle that reinforce the importance of matching the program development stage with the appropriate evaluation strategies, as Scheirer outlines in the life cycle framework:

- “It is essential to design programs and evaluations at the same time so that a program can be implemented in a manner that makes an evaluation feasible.
- “It is also essential to get timely baseline data if before-and-after comparisons are a feature of an evaluation.
- “Evaluation findings need to emerge in a timely manner. Great insights that arrive too late to influence the next steps or federal or state policy have no impact.
- “Evaluation design needs to be flexible enough to adjust to changing features of the program being evaluated. Few programs are implemented as planned, and most end up being less ambitious in scope than originally projected.
- “The evaluation team and the program team need to work together. Whenever tension, a lack of respect, or an inability to work together characterizes a demonstration initiative, it is unusual for useful findings to emerge.
- “Evaluations tend to be stronger when a program has clearly defined and clearly measurable expected outcomes” (Knickman & Hunt, 2008).

Even having articulated these lessons, RWJF has still experienced problems with aligning evaluations with the program life cycle. These have come to light in Program Results Reports (PRR), which RWJF commissions from consultants on every national program after its completion. The reports explain the problem addressed; the strategy used to facilitate change; the results of the program; the evaluation of the program and its findings; lessons for the field; and any postprogram activities. The writers also collect grant making lessons for RWJF staff. These lessons come from program officers, national program directors, evaluators, and project directors. The following illustrates some of the lessons relevant to evaluation and program life cycle that we derived from PRRs completed since 2000. To learn more about any of the programs described briefly below, please see the Publications & Research section of the RWJF website at www.rwjf.org/pr/index.jsp.

¹ Robert Wood Johnson Foundation
Phase I: Program Development

New Health Partnerships: Improving Care by Engaging Patients

New Health Partnerships: Improving Care by Engaging Patients (January 2005 to September 2009) explored whether primary care centers could deliver comprehensive patient- and family-centered self-management support to patients with chronic conditions. Teams at 29 primary care centers participated. The evaluation documented the change processes within systems of clinical care delivery and patient outcomes by examining:

- The extent to which new “change processes” were adopted in these primary care centers, using a self-report instrument for center staff;
- How participation in learning communities facilitated the adoption, sustainability and spread of self-management support services among participating sites, using qualitative telephone interviews; and
- Patient perceptions of care, health behavior and clinical outcomes, using a patient self-report survey form.

Lesson. Start the evaluation at the beginning of the program. The evaluators began working on New Health Partnerships during the planning phase, so the evaluation content was integrated into the program. This enabled evaluators to produce prompt findings that primary care teams could use to improve their work and the national program office could use to plan each phase of the program, as suggested in the Program Development phase of the life cycle framework.

Pursuing Perfection: Raising the Bar for Health Care Performance

Pursuing Perfection: Raising the Bar for Health Care Performance (February 2001 to August 2008) was a national program that supported efforts by seven health care organizations to improve their care processes and patient outcomes. In May 2002, RWJF awarded a $1.5 million, 30-month grant to evaluate Phase 2 of the program.

This evaluation assessed both program development and impact. The evaluation team visited the seven sites multiple times between 2002 and 2005, conducted a series of interviews with staff members and administered a written survey to a broad-based sample of the site personnel. The interview and survey questions focused on several program development issues: (1) how improvement efforts were organized; (2) the role played by top leadership; (3) the degree of progress; (4) barriers to success and factors that helped overcome the barriers; and (5) individual customs and characteristics of each organization.

Lesson. Consider choosing the evaluators of a program early and including them in the program’s design process. The Pursuing Perfection evaluation team began work after the implementation phase was already underway—and after approaches to data collection and measurement were already set. Earlier involvement would have benefited this evaluation.

Phase II: Testing the Program for Causal Efficacy

Hablamos Juntos

Hablamos Juntos (October 2001 to June 2006) was the first national effort to help health care organizations meet the challenge of providing language services and signage to help non-English speakers navigate the health system. The program targeted 10 communities or service areas with new and fast-growing Latino populations.
The evaluation design included case studies describing the intervention at each site and identifying facilitators and barriers to implementing language services. Two waves of patient telephone surveys, focus groups with Latino patients, and other methods helped to assess the impact of these services.

Lesson. Design evaluations with a clear understanding of the developmental state of a field. The evaluator regretted not having invested more time documenting the experiences of sites as they tried to implement something new and untested. The RWJF evaluation officer believed the evaluation’s outcome survey to assess causal impacts was premature because there was no stable intervention to test. She suggested that the case studies involved the sites in a more meaningful and appropriate way and were much more useful to learning, given the early developmental stage of language services. In this example, the evaluation could have placed more emphasis on Phase I methods to aid program development, rather than focusing immediately on causal efficacy.

Phase III: Normal, Continuous Program Delivery

Better Jobs Better Care

Better Jobs Better Care (July 2002 to September 2006) supported changes in long-term-care policy and provider practices designed to improve the quality of care and to reduce high vacancy and turnover rates among the paraprofessionals (nursing assistants, home health aides, and personal care attendants) who provide direct care to older adults. The program included demonstration grants to coalitions in five states and applied research grants to eight teams across the country.

Evaluators conducted a two-part evaluation. First, an implementation evaluation analyzed successes and challenges within the five coalitions, using qualitative data from project work plans and progress reports as well as telephone and in-person interviews with project staff, coalition stakeholders, and state policy experts. These early studies helped inform the field about how direct-care workers viewed their on-going job conditions. Second, an outcome evaluation assessed the impact of provider-level practice changes on recruiting and retaining direct-care workers and other job-related outcomes. Evaluators conducted surveys of clinical managers, direct-care workers and frontline supervisors, and collected direct-care worker employment data from participating providers.

During the course of the program the evaluators made some changes to the original evaluation plan. Prompted by the first round of site visits that showed variation in the degree of intervention implementation, the evaluators added implementation questions to the clinical manager survey in the outcome evaluation. At the next round of site visits, they included questions about the degree of implementation and added in-person site visits to providers, which were not originally planned.

Lesson. Encourage evaluators to include a mid-course correction. According to the evaluator, the mid-course correction was invaluable in ensuring that the evaluation team adequately addressed whether providers had been able to implement their projects. It also helped them realize how important a mixed-method design would be to capture both qualitative and quantitative data in this evaluation of ongoing program and staffing features of long-term care.

Phase IV: Dissemination and Replication of Effective Programs into Widespread Use

Covering Kids & Families

Covering Kids & Families (May 2001 to April 2009) was designed to find, enroll and retain eligible children and adults in government health care coverage programs. Statewide and local
coalitions in 45 states and the District of Columbia participated in the program, with the five remaining states later receiving awards for communications campaigns and meetings. Eighteen local coalitions also participated in the Covering Kids & Families Access Initiative, intended to improve access to health services for children, adults and families covered by public health insurance programs.

Between 2002 and 2008, a team of researchers evaluated Covering Kids & Families to determine the outcomes of RWJF’s investment. The initial evaluation used multiple methods to document grantee strategies and actions; to assess changes that simplified the application processes; and to measure progress in expanding enrollment and retention in Medicaid and the Children’s Health Insurance Program. A separate evaluation of the Covering Kids & Families Access Initiative followed-up on the status of those projects 18 months after RWJF funding ended, to assess sustainability. Researchers also visited Access Initiative projects in five states to develop conclusions and lessons about the experiences of grantees in the initiative.

**Lesson.** Consider doing follow-up evaluations after a program ends, particularly if the funding period is less than 2 years and the program is still evolving as the grant period ends. For example, the follow-up evaluation of the Covering Kids & Families Access Initiative showed that when a grantee organization starts to work with new funders, the project can evolve in new directions and the connection with the original funder’s agenda can grow weaker. As time passes, it may not always be clear that the grantee’s activities can still be viewed as an extension of the original project.

**RWJF Perspectives**

**Program Development**

As the examples above illustrate, RWJF has sometimes incorporated a life cycle approach into the evaluation of its national programs but not always. To encourage greater integration, we involve the Foundation’s evaluation staff in the process of program creation, and we also choose the outside evaluator at the same time we select a national program office and program director. This means the evaluator can be involved in program planning and logistics discussions and the program director can be involved in planning the evaluation. We do not see “independence” in the evaluator as equivalent to a firewall. Evaluations are one of the ways we and our grantees learn. They are not a means to hold grantees accountable, which engenders fear, not learning.

**Testing for Causal Efficacy**

Not all RWJF evaluations include testing for program efficacy or impact. We strive to make the purpose of each evaluation explicit to program grantees, as well as what they should expect and when—and what is required from them during the evaluation process if it does include a rigorous efficacy-testing design. Our goal is to ensure that the roles and responsibilities of both evaluators and grantees are clear.

**Continuous Program Delivery**

Flexibility is key. Programs evolve, and so must evaluations. We negotiate logistics when negotiation is needed. When program experience shows a need to revise the evaluation plan, we work with the evaluator to refocus. There may be binding requirements for grantees or evaluators around specific activities—and we reiterate these as many times as needed—but we are committed to fostering mutual respect among grantees, evaluators, and Foundation staff.
**Dissemination**

Finally, we want our evaluations to be useful to the field and therefore include evaluation reports on our website, along with any articles the evaluators may publish. For RWJF, the purpose of evaluation is learning—for ourselves, for our grantees, for other foundations, for people running similar programs, and for the field at large.

**POSTSCRIPT**

We have just learned from the Center for Effective Philanthropy that grantees generally are favorable to our external evaluation contractors. According to the Center, this stands in contrast to the experience of grantees of other major foundations that use independent evaluators. We believe that the difference may have to do with attention to the process of commissioning evaluations and application of the lessons that have been learned to date.

**Reference**