10

Conceptualizing the Intervention

Alternatives for Evaluating Theories of Change

All the World's a Stage for Theory

In Tony Kushner's Pulitzer Prize–winning play, *Angels in America*, Part Two opens in the Hall of Deputies, the Kremlin, where Aleksii Antedilluvianovich Prelapsarianov, the World's oldest living Bolshevik, speaks with sudden, violent passion, grieving a world without theory:

> *How are we to proceed without Theory? What System of Thought have these Reformers to present to this mad swirling planetary disorganization, to the Inevident Welter of fact, event, phenomenon, calamity? Do they have, as we did, a beautiful Theory, as bold, as Grand, as comprehensive a construct . . . ? You can't imagine, when we first read the Classic Texts, when in the dark vexed night of our ignorance and terror the seed-words sprouted and shoved incomprehension aside, when the incredible bloody vegetable struggled up and through into Red Blooming gave us Praxis, True Praxis, True Theory married to Actual Life. . . . You who live in this Sour Little Age cannot imagine the grandeur of the prospect we gazed upon: like standing atop the highest peak in the mighty Caucasus, and viewing in one all-knowing glance the mountainous, granite order of creation. You cannot imagine it. I weep for you.*
>
> *And what have you to offer now, children of this Theory? What have you to offer in its place? Market Incentives? American Cheeseburgers? Watered-down Bukharinite stopgap makeshift Capitalism? NEPmen! Pygmy children of a gigantic race!*
>
> *Change? Yes, we must change, only show me the Theory, and I will be at the barricades, show me the book of the next Beautiful Theory, and I promise you these blind eyes will see again, just to read it, to devour that text. Show me the words that will reorder the world, or else keep silent.*[1]
>
> —Kushner 1994:13–14

Evaluation and Program Theory

Evaluability Assessment

The idea that evaluation should include conceptualizing and testing a program's theory of change emerged in the 1970s as part of a more general concern about assessing a program's *readiness for evaluation*. The notion was basically this: Before undertaking an evaluation, the program should be clearly conceptualized as some identifiable set of activities that are expected to lead to some identifiable outcomes. The linkage between those activities and outcomes should be both logical and testable. "Evaluability assessment is a systematic process for describing the structure of a program and for analyzing the plausibility and feasibility of achieving objectives; their suitability for in-depth evaluation; and their acceptance to program managers, policymakers, and program operators" (Smith 2005a:136; see also Smith 1989).

One primary outcome of an evaluability assessment is definition of a program's theory. This means specifying the underlying logic (cause and effect relationships) of the program, including what resources and activities are expected to produce what results. An evaluability assessment is also expected to gather various stakeholders' perspectives on the program theory and assess their interest in evaluation. Also assessed are the program's capacity to undertake an evaluation and its readiness for rigorous evaluation (e.g., whether the program's theory is sufficiently well conceptualized and measures of outcomes adequately validated to permit a meaningful summative evaluation).

Evaluability assessment was the evaluator's version of foreplay: getting the program ready for the act itself, the act being evaluation—leading to the climax of producing findings. Or if you find sexual innuendo distracting or inappropriate, consider an agricultural analogy. Evaluability assessment involved tilling the soil before planting the seeds (evaluation questions) that, if properly nourished, would produce an abundant yield (useful findings).

In effect, evaluability assessment puts evaluators in the business of facilitating design of the program in order for it to be evaluated. For already existing programs, this means redesigning the program because the original program model was insufficiently specified to be evaluated. Intended outcomes are often vague or unmeasurable (as discussed in Chapter 7), and how desired outcomes will actually result from the program's activities is often far from clear. As evaluators became involved in working with program people to more clearly specify the program's model (or theory), it became increasingly clear that evaluation was an *up-front activity* not just a back-end activity. That is, traditional planning models laid out some series of steps in which planning comes first, then implementation of the program, and then evaluation, making evaluation a back-end, last-thing-done activity. But to get a program plan or design that could actually be evaluated meant involving evaluators—and evaluative thinking—from the beginning.

Evaluative thinking, then, becomes part of the program design process, including, especially, conceptualizing the program's theory of change: How will what the program does lead to the desired results? Engaging in this work is an example of *process use* (Chapter 5) in which the evaluation has an impact on the program quite apart from producing findings about program effectiveness. The very process of conceptualizing the program's theory of change can have an impact on how the program is implemented, understood, talked about, and improved. The evaluative thinking process has these impacts.

Process Use and Theory of Change

Assisting primary intended users to conceptualize the program's theory of change can have an impact on the program before any evaluative data are gathered about whether the program's theory works. This is an example of the *process use* of evaluation (as opposed to findings use). The very process of conceptualizing the program's theory of change can affect how the program is implemented, understood, talked about, and improved.

This has huge implications for evaluators. It means that evaluators have to be (1) astute at conceptualizing program and policy theories of change and (2) skilled at working with program people, policymakers, and funders to facilitate their articulation of their implicit theories of change. Given the importance of these tasks, it matters a great deal what theory of change frameworks the evaluator can offer. Options for doing theory of change work as part of a utilization-focused evaluation is the subject of this chapter.

Mountaintop Inferences

> *That evil is half-cured whose cause we know.*
>
> —Shakespeare

Causal inferences flash as lightning bolts in stormy controversies. While philosophers of science serve as meteorologists for such storms—describing, categorizing, predicting, and warning, policymakers seek to navigate away from the storms to safe harbors of reasonableness. When studying causality as a graduate student, I marveled at the multitude of mathematical and logical proofs necessary to demonstrate that the world is a complex place (e.g., Nagel 1961; Bunge 1959). In lieu of rhetoric on the topic, I offer a simple Sufi story to introduce this chapter's discussion of the relationship between means and ends, informed and undergirded by theory.

> *The incomparable Mulla Nasrudin was visited by a would-be disciple. The man, after many vicissitudes, arrived at the hut on the mountain where the Mulla (teacher) was sitting. Knowing that every single action of the illuminated Sufi was meaningful, the newcomer asked Nasrudin why he was blowing on his hands. "To warm myself in the cold, of course," Nasrudin replied.*
>
> *Shortly afterward, Nasrudin poured out two bowls of soup, and blew on his own. "Why are you doing that, Master?" asked the disciple. "To cool it, of course," said the teacher.*
>
> *At that point, the disciple left Nasrudin, unable to trust any longer a man who used the same process to cause different effects—heat and cold.*
>
> —Adapted from Shah 1964:79–80

Conceptualizing Interventions

At the simplest level, we can model what the disciple observed as follows:

Hot soup → Blow on hot soup → Cooler soup

Cold hands → Blow on cold hands → Warmer hands

So, what's going on in these two sequences? What's the intervention? The intervention is Nasrudin's breath. The baselines are (1) hot soup and (2) cold hands. The results are (1) cooler soup and (2) warmer hands. We assume Nasrudin's breath temperature to be about the same temperature in each case. Puzzling this out, we can posit the following *intervention theory*: If the object being blown on is warmer than one's breath, then the object will be cooled by the blowing; if the object being blown on is cooler than one's breath,

then the object will be warmed. That's a simple intervention theory. An intervention theory is basically an *if/then* assertion or hypothesis: If we do *x,* then *y* will result.

Now, using this simple sequential logic, let's turn to a program intervention.

Person lacks training needed to get a good job → Provide appropriate training → Trained person gets a good job

This is a simple (and common) program theory. If we train people, then they will get good jobs. It focuses on a single problem: lack of training. It provides a focused intervention: job training. It has a straightforward, measurable outcome: a good job. That's a starting place—and it's the starting place for many policymakers and program designers who want to help poor people get better jobs. Then we start asking deeper questions and surfacing assumptions. Does "training" mean just skill training (how to do keyboarding and data entry), or does it also include "soft skills" (how to get along in the workplace)? What is "appropriate" training? What is a "good" job? At this stage, these aren't measurement questions. We're not asking how we would measure whether or not a person got a good job. We're asking conceptual and values-based questions: Will the kind of training provided lead to the kind of job desired? Is it enough to give the person keyboarding skills? What if the person is a recent immigrant and speaks English poorly? Does the program intervention need to include language training? What if the trainee uses drugs? Does the program need to include drug treatment? What if the poor person is a single mother with young children? Does the program intervention need to include child care? How will the poor person get to training? Will the program intervention have to include transportation support to be effective? Is it enough to provide training, or will there need to be job placement services? And what about the workplace? If the poor person being training is African American and the job opportunities are in companies with mostly white employees, will some kind of support be needed in the workplace to create an environment in which this newly trained person can succeed? As the questioning proceeds, the simple intervention above may morph into the more complicated program intervention as depicted in Exhibit 10.1, which presents the program theory of a real program.

The Jargon Challenge: What Are We Talking About?

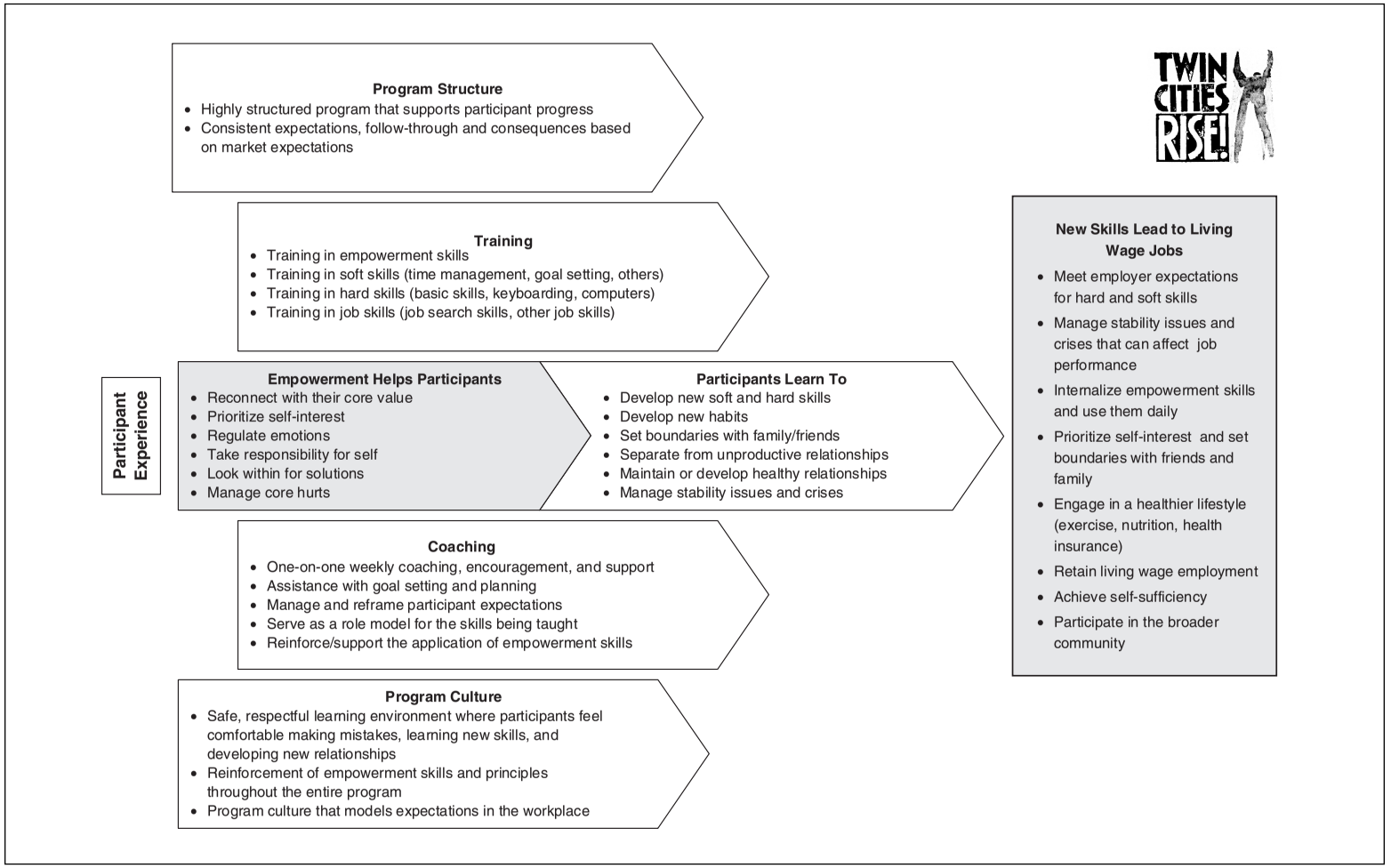
A proliferation of terms has come into use describing how program activities lead to program outcomes. Some of the language emphasizes elucidating the logic of what the program does, so we have logic models, logical frameworks, and intervention logic. Some focus on theory: program theory, theory-based evaluation, theory-driven evaluation, theory of change, theory of action, and intervention theory. Some approaches emphasize linkages: chain of objectives, outcomes mapping, and impact pathway analysis. Three important distinctions are embedded in these different terms.

(1) *Logic modeling versus theory of change.* Does the model simply describe a logical sequence or does it also provide *an explanation* of why that sequence operates as it does? *Specifying the causal mechanisms transforms a logic model into a theory of change.*

A logic model only has to be logical and sequential. The logic of a logic model is partially temporal: It is impossible for an effect or outcome to precede its cause. A logic model expresses a sequence in the sense that one thing leads to another. You crawl before you walk before you run is a descriptive logic model. Crawling precedes walking, which precedes running. It becomes a theory

EXHIBIT 10.1

Theory of Change for an Employment Training Program



of change when you explicitly add the change mechanism. You crawl, and crawling develops the gross-motor skills and body control capabilities that make it possible to walk; you walk, and walking develops the balance, further gross-motor skills, and body control needed to run. *Adding the causal mechanism moves the model from program logic to program theory.*

(2) A second critical distinction involves the source of the model. The terms *program logic* or *program theory* imply that what is being depicted is what people who run the program believe is going on. It is the theory articulated by the program staff, administrators, and funders. In contrast, *theory-driven evaluation* or *theory-based evaluation* typically refers to the program as a test of some larger social science theory. Staff in a faith-based initiative may explain a program by saying it puts participants in touch with their inherent spiritual nature; this would be the program's theory. A social science researcher might look at the same program through the lens of a sociological theory that explains how cohesive groups function to create shared beliefs and norms that determine behavior; that approach would be theory driven. *Theory of change* can be a hybrid of both program theory and social science theory, and often is, as the idea of theory-based evaluation has evolved over the years (Mason and Barnes 2007; Rogers 2007; Weiss 2007) and come to include both "small theories" that are program specific (Leviton 2007; Lipsey 2007c; Layzer 1996) and the larger theories of which a specific program theory is but one manifestation.

(3) A third distinction concerns the *unit of analysis*—or we might say, the *unit of logic,* or the *boundaries of the theory of change.* In elucidating a program model, the term *program* is sometimes a discrete local effort, like a local employment training program. That local program has its own logic model and/or program theory that constitutes a specific intervention. But in large organizations like international development agencies or philanthropic foundations, a program can refer to a collection of interventions made up of several projects and grants. For example, The Atlantic Philanthropies as a philanthropic foundation has a Reconciliation and Human Rights strategic focus that consists of three program areas each with several distinct projects, grants, and intervention strategies, some of which fit together into a cluster. The cluster has its own theory of change distinct from but based on the logic models of individual grants. In such settings, one has to be careful to specify the unit of analysis for the theory of change. The language of intervention logic or intervention theory avoids confusion about what the word "program" means by focusing on a specific intervention which might be one strategy within an umbrella program (that has several interventions) or a strategy that cuts across a number of programs (where the overall intervention is a comprehensive, multifaceted, integrated, and omnibus development strategy). A policy or advocacy process can also be the unit of analysis for which one is developing a logic model (Coffman 2007a; Gardner and Geierstanger 2007; Hendricks-Smith 2007; Kay 2007).

The *theory of action* language comes from action research and organizational development traditions where the focus is typically on some specific solution to a specific problem (the "action"); that action may not be a full-scale intervention, program, policy, or theory—but an action taken within some concrete time period for some specific purpose. Doing something to reduce the dropout problem in a program would involve some theory of action. The theory of action tradition places particular emphasis on distinguishing "espoused theory" (how practitioners explain what they are attempting to do) from "theory-in-use" (what their

Program Theory and Logic Model Babel

Confusion reigns in the language describing how program activities lead to program outcomes:

* logic model * logical framework * program logic * program model * intervention logic * intervention model * chain of objectives * outcomes map * impact pathway analysis * program theory * theory-based evaluation * theory-driven evaluation * theory of change * theory of action * intervention theory *

Which term is best? That best designation is the one that makes the most sense to primary intended uses—the term they resonate to and has meaning within their context given the intended uses of the evaluation.

behavior reveals about what actually guides what they do). The major evaluative thrust of the theory of action framework is helping practitioners examine, reflect on, and deal with the discrepancies between their espoused theory and their theory-in-use (Argyris 1993, 1974; Schön 1987, 1983; Argyris and Schön 1978, 1974). "People do not always behave congruently with their beliefs, values, and attitudes (all part of espoused theories). . . . Although people do not behave congruently with their espoused theories, they do behave congruently with their theories-in-use" (Argyris 1982:85).

In this conundrum of dissonance between stated belief and actual practice lies a golden opportunity for reality testing: the heart of evaluation. Sociologist W. I. Thomas posited in what has become known as *Thomas' Theorem* that *what is perceived as real is real in its consequences.* Espoused theories are what practitioners perceive to be real. Those espoused theories, often implicit and only espoused when asked for and coached into the open, have real consequences for what practitioners do. Elucidating the theory of change held by primary users can help them be more deliberative about what they do and more willing to put their beliefs and assumptions to an empirical test through evaluation. In short, the user-focused approach challenges decision makers, program staff, funders, and other users to engage in reality testing, that is, to test whether what they believe to be true (their espoused theory of action) is what actually occurs (theory-in-use).

Which Term Is Best?

Given all this diversity of and confusion in language, which term is best? Logic model? Theory of change? Intervention model? From a utilization-focused evaluation perspective, *that label is best that makes the most sense to primary intended uses*—the term they resonate to and has meaning within their context. Here are some examples of how contexts vary.

In international settings, logical frameworks or "Logframes" have a long history of use by government aid agencies (Norwegian Agency for Development Cooperation 1999). United Way of America has promoted logic models as a way for nonprofit agencies to present funding proposals (United Way 1996). Large-scale community development initiatives have adopted the "theory of change" language due in large part to an influential article by Carol Weiss (1995) widely disseminated by The Aspen Institute (Connell et al. 1995). *Program theory* has been a target for "advancement" among evaluators for more than two decades (Bickman 1994, 1990). *Program logic* and *program theory* are familiar terms in Australia and New Zealand due to the influential works of Bryan Lenne (1987), Sue Funnell (2005, 2000, 1997), and Patricia Rogers (2008, 2005b, 2005c, 2003, 2000a, 2000b). Theory-driven evaluation has been widely promoted by Huey-Tsyh Chen (2005a, 2005b, 2004, 1990) and, like its cousin, theory-based evaluation (Weiss 2007, 2000, 1997; Birckmayer and Weiss 2000) is a label that plays well in academic settings,

theory being much revered in universities. Stewart Donaldson (2007), Director of the Institute of Organizational and Program Evaluation Research at Claremont Graduate University, argues that systematic attention to program theory in evaluation elevates evaluation to the status and prestige of *science*, what he calls *Program Theory-Driven Evaluation Science*. From a utilization-focused perspective, the choice of label depends on the purpose of the conceptual work and the preferences of primary intended users. As Rogers (2005c) has observed,

> Program logic is sometimes used interchangeably with program theory. . . . In many cases, the choice of term is based on local responses to the words *theory* and *logic* (each of which can be seen as unpalatable) and on the terms used in the specific texts used by the evaluators. (P. 339)

The challenge, then, is to use terms that have meaning within a particular context and tradition. We'll now look at some of these approaches more closely.

The Logic Model Option in Evaluation: Constructing a Means-Ends Hierarchy

> *Causation. The relation between mosquitos and mosquito bites.*
>
> —Michael Scriven (1991b:77)

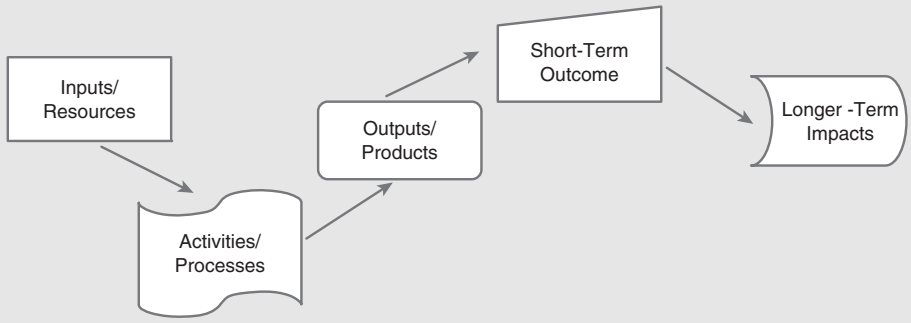
A theory links means and ends. The construction of a means-ends hierarchy for a program constitutes a comprehensive description of the program's model. For example, in his classic work on evaluation, Suchman (1967) recommended building a *chain of objectives* by trichotomizing objectives into immediate, intermediate, and ultimate goals. The linkages between these levels make up a continuous series of actions wherein immediate objectives (focused on implementation) logically precede intermediate goals (short-term outcomes) and therefore must be accomplished before higher-level goals (long-term impacts). Any given objective in the chain is the outcome of the successful attainment of the preceding objective and, in turn, is a precondition to attainment of the next higher objective.

> Immediate goals refer to the results of the specific act with which one is momentarily concerned, such as the formation of an obesity club; the intermediate goals push ahead toward the accomplishment of the specific act, such as the actual reduction in weight of club members; the ultimate goal then examines the effect of achieving the intermediate goal upon the health status of the members, such as reduction in the incidence of heart disease. (Suchman 1967:51–52)

The means-ends hierarchy for a program often has many more than three links. In Chapter 7, I presented the mission statement, goals, and objectives of the Minnesota Comprehensive Epilepsy Program. One of the goals was to conduct high-quality research on epilepsy. Exhibit 10.2 presents the chain of objectives for that goal.

The full chain of objectives that links inputs to activities, activities to immediate outputs, immediate outputs to intermediate outcomes, and intermediate outcomes to ultimate goals constitutes a program's logical model. Any particular paired linkage in the theory displays an action and reaction: a hypothesized cause and effect. As one constructs a hierarchical/sequential model, it becomes clear that there is only a relative distinction between ends and means: "Any end or goal can be seen as a means to another goal, [and] one is free to enter the 'hierarchy of means and ends' at any point" (Perrow 1968:307). *In utilization-focused evaluation, the decision about where to enter the means-ends hierarchy for a particular evaluation is made on the basis of what information would be most*

BASIC LOGIC MODEL



Logic Modeling Resources

Centers for Disease Control Program Evaluation Resources
http://www.cdc.gov/healthyyouth/evaluation/resources.htm#4

European AID. 2005. *Evaluation Tools.* Brussels: European Commission.
http://ec.europa.eu/europeaid/evaluation/methodology/tools/too_en.htm

Frechtling. 2007. *Logic Modeling Methods in Program Evaluation.*

Kellogg Foundation. 2001. *Logic Model Development Guide: Logic Models to Bring Together Planning, Evaluation & Action.*
http://www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf

Rogers. 2005b. *Logic Model.*

United Way. 1996. *Measuring Program Outcomes: A Practical Approach.*
http://national.unitedway.org/outcomes/resources/mpo/model.cfm

University of Wisconsin Cooperative Extension. 2007. *Program Development and Evaluation.*
http://www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html

useful to the primary intended evaluation users. In other words, a formative evaluation might focus on the connection between inputs and activities (an implementation evaluation) and not devote resources to measuring outcomes higher up in the hierarchy until implementation was ensured. Elucidating the entire hierarchy does not incur an obligation to evaluate every linkage in the hierarchy. The means-ends hierarchy displays a series of choices for more focused evaluations while also establishing a context for such narrow efforts. Suchman (1967:55) used the example of a health education campaign to show how a means-ends hierarchy can be stated in terms of a series of measures or evaluation findings. See Exhibit 10.3.

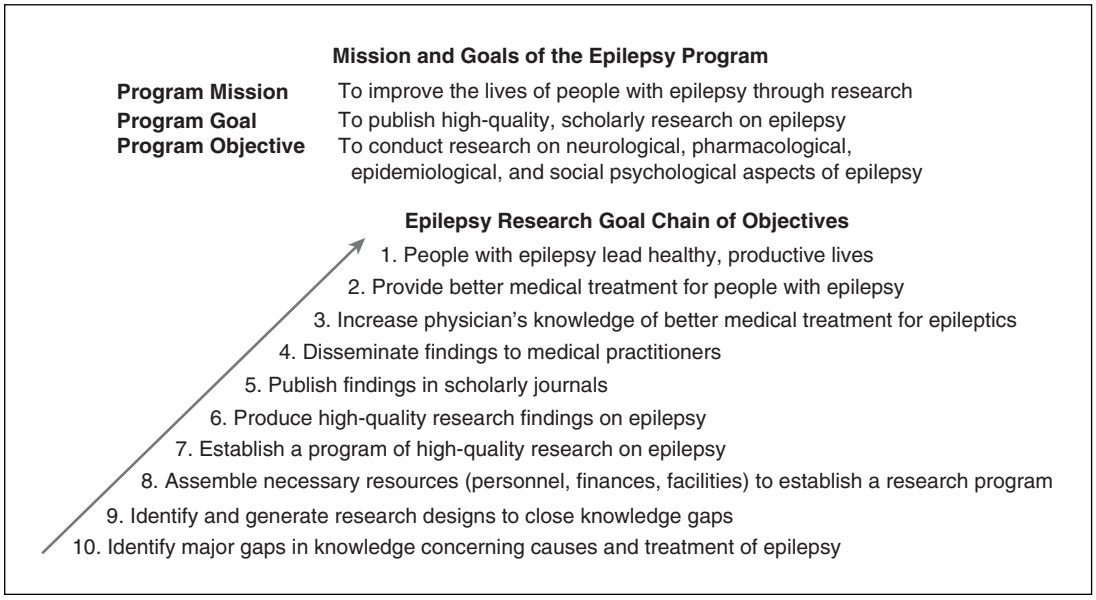
Testing a Logic Model: A Policy Implementation Example

Let me offer a simple example of testing a logic model. A State Department of Energy allocated conservation funds through 10 regional districts. An evaluation was commissioned by the department to assess the impact of local involvement in priority setting. State and regional officials articulated the following *fair and equitable logic model* of decision making:

1. State officials establish funding targets for energy proposals from districts and rules for submitting proposals.

EXHIBIT 10.2

Epilepsy Program Logic Model

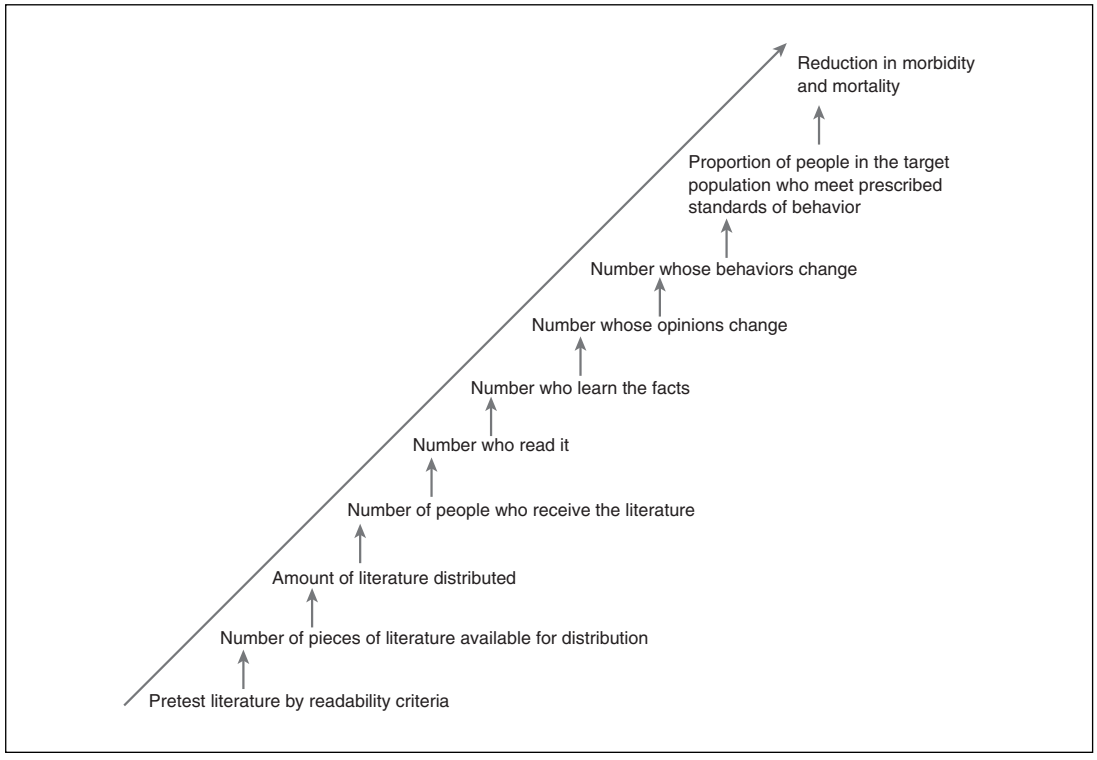


2. District advisory groups assess each district's energy needs with broad citizen input and involvement.
3. District advisory groups develop funding proposals, based on the needs assessments that meet their district's state target and follow the state's rules.
4. The state approves the budgets based on the merit of the proposals within the guidelines, rules, and targets provided.
5. Expected results: (a) Approved funds equal original targets and (b) everyone perceives the funding as fair and equitable.

In short, the espoused logic model was that decisions would be made fairly based on explicit and transparent procedures, guidelines, and rules. The data showed this to be the case in only 6 of the 10 districts. In the other 4 districts, proposals from the districts exceeded the assigned target amounts by 30 percent to 55 percent; for example, one district, assigned a target of $100 million by the state, submitted a proposal for $140 million (despite a "rule" that said proposals could not exceed targets). Moreover, the final, approved budgets exceeded the original targets by 20 percent to 40 percent. The district with a target of $100 million and a proposal for $140 million received $120 million. Four of the districts, then, were not engaged in a by-the-book equitable process; rather, their process was negotiated, personal, and political—and subsequently perceived as unfair. Needless to say, when these data were presented, the six districts that followed the guidelines and played the funding game by what they thought were uniform rules—the districts whose proposals equaled their assigned targets—were outraged. Testing the espoused theory of fairness and uniform

EXHIBIT 10.3

Logic Model Hierarchy of Evaluation Measures for a Health Education Campaign



SOURCE: Adapted from Suchman 1967:55.

rules revealed that the reality (theory-in-use) in four districts did not match the espoused ideal in ways that had significant consequences for all concerned.

This is a simple, commonsense example of testing a logic model at the policy level. Nothing elegant. No academic trappings. The espoused model is a straightforward articulation of what all agreed was supposed to happen in the process to achieve desired outcomes. The linkages between processes and outcomes are made explicit. Evaluative data then revealed what actually happened—and where what actually happened departed from what was supposed to happen and the consequences of that discrepancy. At the district level, each district would have its own model of how funds were to be used and what those funds were supposed to accomplish.

Three Approaches to Program Theory

A logic model only has to be logical and sequential. Adding specification of the causal mechanism moves the model from program

The Logical Framework Matrix

A logical framework (Logframe) is a matrix that specifies each step in the chain of objectives and requires for each step that a target result be specified, that the data source be specified for measuring the result, and that any critical assumptions be stated. The *Logical Framework Approach* (IFAD 2002; NORAD 1999; Sartorius 1996, 1991) offers a format for connecting levels of impact with evidence. Used widely by international development agencies as a comprehensive map in designing projects, the framework begins by requiring specification of the overall goal and purposes of the project. Short-term outputs are linked logically to those purposes, and activities are identified that are expected to produce the outputs. Careful attention must be paid to the language of this model because what the logical framework calls a goal is what other models more commonly call mission; and "purposes" in this approach are similar to objectives or outcomes; outputs are short-term, end-of-project deliverables. For every goal, purpose, output, and activity, the framework requires specification of objectively verifiable indicators, means of verification (types of data), and important assumptions about the linkage between activities and outputs, outputs to purposes, and purposes to goals. In their very practical book on *RealWorld Evaluation*, Bamberger, Rugh, and Mabry (2006) discuss how to integrate logical framework and program theory approaches with excellent examples from their international development experiences.

logic to program theory. In this section, we will look at three major approaches to developing program theory for evaluation use:

1. *The deductive approach*—drawing on scholarly theories from the academic literature
2. *The inductive approach*—doing fieldwork on a program to generate grounded theory, for example, as part of an evaluability assessment process
3. *The user-focused approach*—working with intended users to extract and make explicit their implicit theory of action

The deductive approach draws on dominant theoretical traditions in specific scholarly disciplines to construct models of the relationship between program treatments and outcomes. For example, an evaluation of whether a graduate school teaches students to think critically could be based on the theoretical perspective of *a phenomenography of adult critical reflection*, as articulated by the Distinguished Professor of Education Stephen Brookfield (1994), an approach that emphasizes the visceral and emotional dimensions of critical thought as opposed to purely intellectual, cognitive, and skills emphases. Illustrations of the deductive approach to evaluation are chronicled in Chen (2004), Rossi and Freeman (1993), Lipsey and Pollard (1989), and Chen and Rossi (1989, 1987). Testing social science theories may be a by-product of an evaluation in which the primary purpose is knowledge generation (see Chapter 4). However, the temptation in the deductive approach is to make the study more research than evaluation, that is, to let the scholarly contribution and theory testing take over the evaluation, making it useful academically but not necessarily useful to practitioners and policymakers.

The inductive approach involves the evaluator in doing fieldwork to generate program theory. Staying with the example of evaluating whether graduate students learn to think critically, the inductive approach would involve assessing student work, observing students in class, and interviewing students and professors to determine what model of education undergirds efforts to impart critical thinking skills. Such an effort could be done as an evaluation study unto itself, for example, as part of an early evaluability assessment process, or it could be done in conjunction with a deductive

effort based on a literature review. The product of the inductive approach, and therefore a major product of the evaluation, would be an empirically derived theoretical model of the relationship between program activities and outcomes framed in terms of important contextual factors.

User-Focused Theory of Change Approach

In neither the deductive nor inductive approach to program theory does the evaluator have to engage with key stakeholders. That changes in *the user-focused approach* in which the evaluator's task is to facilitate intended users, especially program personnel, to articulate their operating theory. Continuing with the critical thinking example, this would mean bringing together students and professors to make explicit their educational assumptions and generate a program theory model that could then be tested as part of the evaluation.

Facilitating the program theory articulation process involves working with those who are knowledgeable about the program to construct a flowchart of what happens from the time a person enters a program to the time they leave. How are people recruited or selected into the program? What's their baseline situation when they enter? How are they oriented to the program once they enter? What are the early activities they engage in? How much are they expected to participate? What are they supposed to be doing, learning, acquiring, or changing during those early activities? During later activities? What's the sequence or stages of participants' experiences? What happens as the program approaches the end? What changes should have occurred in participants by the time they reach the end of the program? What mechanisms explain why these changes take place? To what extent are these changes expected to be sustained? What additional outcomes are expected to occur after leaving the program (e.g., keeping a job or staying off of drugs).

The primary focus of utilization-focused evaluation is testing practitioner theories about why they do what they do and what they think results from what they do. Utilization-focused evaluation involves primary intended users in specifying the program's theory and in deciding how much attention to give to testing the theory generated, including how much to draw on social science theory as a framework for the evaluation (Patton 1989).

A Menu of Theory-Based Approaches

Each of the three approaches to program theory—deductive, inductive, and user-focused—has advantages and disadvantages. These are reviewed in Menu 10.1. The strategic calculations a utilization-focused evaluator must make include determining how useful it will be to spend time and effort elucidating a theory of change (or more than one where different perspectives exist, which is common); how to keep theory generation from becoming esoteric and overly academic; how formal to be in the process; and what combinations of the three approaches, or relative emphasis, should be attempted. Factors to consider in making these calculations will be clearer after some more examples, which follow.

Theory of Change and Evaluation: Getting at Causal Mechanisms and Assumptions

The purpose of thoroughly delineating a program's theory of change is to assist practitioners in making explicit their assumptions about the linkages between inputs, activities, immediate outputs,

MENU 10.1

Approaches to Generating Program Theory

Approach	*Potential Advantages*	*Potential Disadvantages*	*Pitfalls to Avoid*
User-focused approach: working with intended users to extract and specify their implicit theory of action to make it explicit action	Intended users understand the theory of action Intended users own the theory of action	As users struggle to articulate their theory, they may be defensive Formal explicit model may not reflect program realties	Don't oversimplify to the point of esoteric meaninglessness in an effort to manage conflict or varying perceptions Don't force articulation of a single theory. Different users may well have different theories of action
Inductive approach: doing fieldwork on a program to generate grounded theory	Theory grounded in real-world practice High relevance because theory is generated from actual program activities and observed outcomes Can focus an evaluability assessment effort	Fieldwork takes time and resources for evaluation and program Likely that different program people operate with different theories in large, multilevel, or complex programs	Don't force a single theory or model on the program where multiple theories of action are operating Don't let generating theory take on a life of its own and become a higher priority than generating useful results
Deductive approach: drawing on scholarly theories from the academic literature	Draws on existing knowledge and literature High academic credibility Connects to larger issues	May not be relevant to specific program May feel esoteric to practitioners Literature search takes time and resources	Don't force program into a theory pigeonhole Don't let theory testing become higher priority than generating useful results
Combining approaches	Uses strengths of each approach Provides diverse perspectives	Costly, time-consuming May lead to conflicting results	Don't let one approach trump another; treat each on its merits, fairly and in balance

Expert Advice from Carol Weiss on Theory-Based Evaluation

- When social science provides theory and concepts that ground and support local formulations, it can be of great evaluative value. The evaluator should bring her knowledge of the social science literature to bear on the evaluation at hand.
- When a number of different assumptions are jostling for priority, a theory-based evaluation is wise to include multiple theories. . . . . But the more theories that are tracked, the more complex and expensive evaluation. Choices have to be made.
- Select theories to test on the following criteria:
    - The first criterion is the beliefs of the people associated with the program, primarily the designers and developers who planned the program, the administrators who manage it, and the practitioners who carry it out on a daily basis. Also important may be the beliefs of the sponsors whose money funds the program and the clients who receive the services of the program. What do these groups assume are the pathways to good outcomes?
    - A second criterion is plausibility. . . . Can the program actually do the things the theory assumes, and will the clients be likely to respond in the expected fashion?
    - A third criterion is lack of knowledge in the program field. This allows the evaluation to contribute knowledge to the field.
    - A final criterion for choosing which theories to examine in a theory-based evaluation is the centrality of the theory to the program. Some theories are so essential to the operation of the program that no matter what else happens, the program's success hinges on the viability of this particular theory.

SOURCE: Weiss (2000:38–41).

intermediate outcomes, and ultimate goals. Suchman (1967) called beliefs about cause-effect relationships the program's *validity assumptions*. For example, many education programs are built on the validity assumptions that (1) new information leads to attitude change and (2) attitude change affects behavior. These assumptions are testable. Does new knowledge change attitudes? Do changed attitudes lead to changed behaviors? Carol Weiss (2000) has commented on the widespread nature of these assumptions:

> Many programs seem to assume that providing information to program participants will lead to a change in their knowledge, and increased knowledge will lead to positive change in behavior. This theory is the basis for a wide range of programs, including those that aim to reduce the use of drugs, prevent unwanted pregnancies, improve patients' adherence to medical regimens, and so forth. Program people assume that if you tell participants about the evil effects of illegal drugs, the difficult long-term consequences of unwed pregnancies, and the benefits of complying with physician orders, they will become more conscious of consequences, think more carefully before embarking on dangerous courses of action, and eventually behave in more socially acceptable ways.
>
> The theory seems commonsensical. Social scientists—and many program people—know that it is too simplistic. Much research and evaluation has cast doubt on its universal applicability. . . . So much effort is expended in providing information in an attempt to change behavior that careful investigation of this theory is warranted. (Pp. 40–41)

Knowing this, when an evaluator encounters a program theory that posits that information will produce knowledge change, and knowledge change will produce behavior change, it is appropriate to bring to the attention of those involved the

substantial evidence that this model doesn't work. In being active-reactive-interactive-adaptive when working with primary intended users, the evaluator can and should bring social science and evaluation knowledge to the attention of those with whom they're working.

Consider this example. The World Bank provided major funding for a program in Bangladesh aimed at improving maternal and child health and nutrition. The theory of change was the classic one we are reviewing here: Information leads to knowledge change, knowledge change leads to practice change. It didn't work. Women in extreme poverty did not have the resources to follow the desired behaviors, even if they were inclined to do so (in other words, they could not afford the recommended foods). Moreover, they live in a social system where what they eat is heavily influenced, even determined, by their mothers-in-law and husbands. The World Bank commissioned an impact evaluation of the project which documented this substantial "knowledge-practice gap" and found that the program was ineffective in closing the gap. "All forms of knowledge transmitted by the project suffer from a knowledge-practice gap, so attention needs to be paid to both the resource constraints that create this gap and transmitting knowledge to other key actors: mothers-in-law and husbands" (World Bank 2005:43).

The larger question is, "Why are those designing interventions aimed at women in extreme poverty still operating on a theory of change that has been discredited time and time again?" We will return to this question later in this chapter, when we discuss bringing systems thinking to bear on program theory. For the moment, I want to use this example to introduce the critical role evaluators play in helping surface and then test a program's causal assumptions.

Identifying Critical Assumptions

Validity assumptions are the presumed causal mechanisms that connect steps in a logic model turning it into a theory of change. The proposition that gaining knowledge will lead to behavior change is undergirded by a validity assumption, namely, that the reason people aren't behaving in the desired manner is because they lack knowledge about what to do. Poor women in Bangladesh don't eat the right foods when they are pregnant because they don't know enough about proper nutrition. Teach them about proper nutrition and they will eat the right foods. It turned out that they gained the knowledge but didn't change their behavior. The validity assumption proved false, or at least insufficient. Knowledge of nutrition may be *a necessary but not sufficient condition* for proper eating.

As validity assumptions are articulated in a means-ends hierarchy, the evaluator can work with intended users to focus the evaluation on those critical linkages where information is most needed at that particular point in the life of the program. It is seldom possible or useful to test all the validity assumptions or evaluate all the means-ends linkages in a program's theory of action. The focus should be on testing the validity of critical assumptions. *In a utilization-focused evaluation, the evaluator works with the primary intended users to identify the critical validity assumptions where reduction of uncertainty about causal linkages could make the most difference.*

While the evaluators can and should bring their own knowledge of social science to bear in interactions with primary intended users, the evaluator's beliefs about critical assumptions is ultimately less important than what staff and decision makers believe. An evaluator can often have greater impact by helping program staff and decision makers empirically test their own causal hypotheses than by telling

Theory of Change

A Theory of Change defines all building blocks required to bring about a given long-term goal. This set of connected building blocks—interchangeably referred to as outcomes, results, accomplishments, or preconditions—is depicted on a map known as a pathway of change/change framework, which is a graphic representation of the change process.

Built around the pathway of change, a Theory of Change describes the types of interventions (a single program or a comprehensive community initiative) that bring about the outcomes depicted in the pathway of a change map. Each outcome in the pathway of change is tied to an intervention, revealing the often complex web of activity that is required to bring about change.

A Theory of Change would not be complete without an articulation of the assumptions that stakeholders use to explain the change process represented by the change framework. Assumptions explain both the connections between early, intermediate, and long-term outcomes and the expectations about how and why proposed interventions will bring them about. Often, assumptions are supported by research, strengthening the case to be made about the plausibility of theory and the likelihood that stated goals will be accomplished.

Stakeholders value theories of change as part of program planning and evaluation because they create a commonly understood vision of the long-term goals, how they will be reached, and what will be used to measure progress along the way.

SOURCE: *ActKnowledge* and the Aspen Institute Roundtable on Community Change (2007) www.theoryofchange.org.

See also Anderson (2005).

them such causal hypotheses are nonsense. This means working with them where they are. So despite my conviction that knowledge change alone seldom produces behavior change, I still find myself helping young program staff rediscover that lesson for themselves. Not only does the wheel have to be re-created from time to time, its efficacy has to be restudied and reevaluated. The evaluator's *certain belief* that square wheels are less efficacious than round ones may have little impact on those who believe that square wheels are effective. The evaluator's task is to delineate the belief in the square wheel, share other research on square wheels when available, and if they remain committed to a square wheel design, assist the true believers in designing an evaluation that will permit them to *test for themselves* how well it works.

I hasten to add that this does not mean that the evaluator is passive. In the active-reactive-interactive-adaptive process of negotiating the evaluation's focus and design, the evaluation facilitator can suggest alternative assumptions and theories to test, but first priority goes to evaluation of validity assumptions held by primary intended users.

Filling in the Conceptual Gaps and Testing the Reasonableness of Program Theories

Helping stakeholders examine conceptual gaps in their theory of change is another task in building program theory and making it evaluatable. In critiquing a famous prison reform experiment, Rutman (1977) has argued that the idea of using prison guards as counselors to inmates ought never have been evaluated (Ward, Kassebaum, and Wilner 1971) because, *on the face of it, the idea is nonsense*. Why would anyone ever believe that prison guards could also be inmate counselors? But clearly, whether they should have or

not, some people did believe that the program would work. Without reaching an evaluation conclusion prior to gathering data, the evaluator can begin by filling in the conceptual gaps in this program theory so that critical validity assumptions can be identified and examined. For example, is some kind of screening and selection part of the design so that the refined theory is that only certain kinds of guards with certain characteristics and in certain roles can serve as counselors to particular kinds of inmates? And what kind of training program for guards is planned? In what ways are guards supposed to be changed during such training? How will changed guard behavior be monitored and rewarded? The first critical assumptions to be evaluated may be whether prison guards can be recruited and trained to exhibit desired counselor attitudes and behaviors. Whether prison guards can learn and practice human relations skills can be evaluated without ever implementing a full-blown program.

Filling in the gaps in the program's theory of change goes to the heart of the implementation question: What series of activities must take place before there is reason even to hope that the desired outcomes will result? I once reviewed a logic model for an after-school program that would provide crafts, arts, and sports activities for middle-school students once a week for 2 hours for one semester—a total of 30 contact hours. The expected outcome was "increased self-esteem." On the face of it, is this a reasonable outcome? What are the causal mechanisms that link 30 hours of after-school group activities to increased self-esteem. Or consider a "dress-for-success" program that provided appropriate dress clothes for poor people to wear to job interviews. The program's stated outcome was the rather modest goal that the appearance of job applicants would make a positive impression. To justify funding the program, the funder wanted the program to change the outcome to getting a job. Now the clothes might help in that regard, but would, at best, be a minor factor. The "dress-for-success" program appropriately resisted being evaluated on the criterion of whether those to whom they provided clothes got a job.

The logic of a chain of objectives is that if activities and objectives lower in the means-ends hierarchy will not be achieved or cannot be implemented, then evaluation of ultimate outcomes is problematic.

> There are only two ways one can move up the scale of objectives in an evaluation: (a) by proving the intervening assumptions through research, that is, changing an assumption to a fact, or (b) by assuming their validity without full research proof. When the former is possible, we can then interpret our success in meeting a lower-level objective as automatic progress toward a higher one. (Suchman 1967:57)

This has important implications for what kind of follow-up is needed in an evaluation to determine impact. Research shows that children immunized against polio do not get polio. The causal connection between the immunization and immunity against polio is established. Therefore, the evaluation can stop at determining that children have been immunized and confidently calculate how many cases of polio have been prevented based on epidemiological research. The evaluation design does not have to include follow-up to determine whether immunized children get polio. That question has been settled by research.

One important reason for testing critical validity assumptions is that some findings are counterintuitive. The Bangladesh maternal and child nutrition program provides an excellent example of the interface between research and evaluation. The program theory posited that proper nutrition for women during pregnancy would reduce the incidence of low birth weight babies. It

seems commonsensical that the proper focus for maternal health would be on nutrition *during the pregnancy*. The evaluation findings questioned this assumption and added to the interpretation results from research showing that the pregnant women's *pre-pregnancy weight* was more predictive of babies' birth weight than weight gain during pregnancy. The evaluation concluded,

> Supplementary feeding for pregnant women appears to be a flawed approach on two grounds: (1) the pregnancy weight gain achieved is mostly too small to have a noticeable impact on birth weight; and (2) it is pre-pregnancy weight that evidence suggests to be the most important determinant of birth weight. . . . The fact that it is pre-pregnancy weight that matters suggests that a different approach altogether ought perhaps be considered, such as school feeding programs or targeting adolescent females in poorer areas. (World Bank 2005:43)

However, this evaluation finding appears to come with its own assumption, namely, that there are not sufficient resources to provide needed food and nutritional supplements to impoverished women *both before and during pregnancy*. By framing the evaluation conclusion as a choice between providing nutrition before pregnancy *or* during pregnancy, the evaluator has limited the policy and programming options. This illustrates how evaluators' own theories of change and assumptions come into play and need to be made explicit and questioned.

Using the Theory of Change to Focus the Evaluation: The New School Case

Once an espoused theory of change is delineated, the issue of evaluation focus remains. This involves more than mechanically evaluating lower-order validity assumptions and then moving up the hierarchy. Not all linkages in the hierarchy are amenable to testing; different causal linkages require different resources for evaluation; data-gathering strategies vary for different objectives. In a summative evaluation, the focus will be on outcomes attainment and causal attribution. For formative evaluation, the most important factor is determining what information would be most useful at a particular point in time. This means identifying those targets of opportunity where additional information could make a difference to the direction of incremental, problem-oriented, program decision making. Having information about and answers to those select questions can make a difference in what is done in the program. Here's an example.

The New School of Behavioral Studies in Education, University of North Dakota, was established to support educational innovations that emphasized individualized instruction, better teacher-pupil relationships, and an interdisciplinary curriculum. The New School established a master's degree, teaching-intern program in which interns replaced teachers without degrees so that the latter could return to the university to complete their baccalaureates. The cooperating school districts released those teachers without degrees who volunteered to return to college and accepted the master's degree interns in their place. Over 4 years, the New School placed 293 interns in 48 school districts and 75 elementary schools, both public and parochial. The school districts that cooperated with the New School in the intern program contained nearly one third of the state's elementary school children.

The Dean of the New School formed a task force of teachers, professors, students, parents, and administrators to evaluate the program. We constructed the theory of change shown in Exhibit 10.4. The objectives stated in the first column are a far cry from being clear, specific, and measurable, but they were quite adequate for discussions aimed at focusing the evaluation question. The second

EXHIBIT 10.4

The New School Theory of Action: A Hierarchy of Objectives, Validity Assumption Linkages, and Evaluation Criteria

Hierarchy of Objectives	*Causal Assumption Linkages*	*Evaluative Criteria*
I. Ultimate Objectives		
1. Prepare children to live full, rich, satisfying lives as adults.	Children whose affective and cognitive needs are met will lead fuller, richer, more satisfying lives as adults.	1. Longitudinal measures of child and adult satisfaction, happiness, and success.
2. Meet the affective and cognitive needs of individual children in North Dakota and the United States.	More open classrooms will better meet the affective and cognitive needs of individual children.	2. Measures of student affective and cognitive growth in open and traditional schools.
3. Facilitate and legitimize the establishment and maintenance of a larger number of more open classrooms in North Dakota and the United States.	Parents and administrators will favor and expand open education once they have experienced it firsthand.	3. Measures of increases in the number of open classrooms in North Dakota and the United States over time and measures of the influence of the New School on the number of open classrooms.
II. Intermediate Objectives		
4. Provide parents and administrators in North Dakota with a firsthand demonstration of the advantages of open education.		4. Measures of parent and administrator attitudes toward New School classrooms and open education, and measures and analysis of the factors affecting their attitudes.

Hierarchy of Objectives	*Causal Assumption Linkages*	*Evaluative Criteria*
	Teachers who have experienced the New School summer program can and will conduct open classrooms during the following intern year that are visible to local parents and administrators.	
5. Provide teachers and teachers-in-training with a one-year classroom experience in conducting an open classroom.		5. Measures of the degree of openness of New School teaching intern classrooms and the factors affecting the degree of openness of these classrooms.
III. Immediate Objectives	Teachers who have experienced the summer program can and will conduct open classrooms.	
6. Provides teachers and teachers-in-training with a summer program in how to conduct open classrooms.		6. Measures of teacher attitudes, teacher understanding, and teacher competency before and after the New School Program.
	To learn about open education, it is best to experience it. Teachers teach the way they are taught.	
7. Provide teachers and teachers-in-training with a personalized and individualized learning experience in an open learning environment.		7. Measures of the degree to which the New School training program is individualized and personalized, and measures of the cognitive and affective growth of teachers in the New School Program.

NOTE: The validity assumptions (middle column) link objectives (left column). Arrows indicate to which objectives the assumptions apply.

column lists validity assumptions underlying each linkage in the theory of action. The third column shows the measures that could be used to evaluate objectives at any level in the hierarchy. When the Evaluation Task Force discussed the program theory, members decided they already had sufficient contact with the summer program to assess the degree to which immediate objectives were being met. With regard to the ultimate objectives, the task force members thought it was premature to evaluate the ultimate outcomes of open education (Objectives 1 and 2), nor could they do much with information about the growth of the open education movement (Objective 3). However, a number of critical uncertainties surfaced at the level of intermediate objectives. Once students left the summer program for the 1-year internships, program staff members were unable to carefully and regularly monitor intern classrooms. They didn't know what variations existed in the openness of the classrooms, nor did they have reliable information about how local parents and administrators were reacting to intern classrooms. Those objectives were prime targets for formative evaluation focusing on three questions: (1) To what extent are summer trainees conducting open classrooms during the regular year? (2) What factors are related to variations in classroom "openness"? (3) What is the relationship between variations in classroom openness and parent/administrator reactions to intern classrooms?

At the onset, nothing precluded evaluation at any of the seven levels in the hierarchy of objectives. There was serious discussion of all levels and alternative foci. In terms of the educational literature, the issue of the outcomes of open education could be considered most important; in terms of university operations, the summer program would have been the appropriate focus; but in terms of the information needs of the primary decision makers and primary intended users on the task force, evaluation of the intermediate objectives had the highest potential for generating useful, formative information.

Theory Informing Practice, Practice Informing Theory

Comparing Program Theories

Much evaluation involves comparing different programs to determine which is more effective or efficient. Evaluations can be designed to compare the effectiveness of two or more programs with the same goal, but if those goals do not bear the same importance in the two programs' theories, the comparisons may be misleading. As part of undertaking a comparative evaluation, it is useful to compare program theories in order to understand the extent to which apparently identical or similarly labeled programs are in fact comparable.

Programs with different intended outcomes cannot be fairly compared to each other on a same outcomes basis. Teacher centers established to support staff development and resource support for school teachers provide an example. The U.S. Office of Education proposed that teacher centers be evaluated according to a single set of universal outcomes. But the evaluation found that teacher centers throughout the country varied substantially in both program activities *and* goals. Exhibit 10.5 describes three types of teacher centers, behavioral, humanistic, and developmental, and summarizes the variations among these types of centers.

Different teacher centers were trying to accomplish different outcomes, so to determine which one was *most effective* became problematic because they were trying to do different things. Evaluation could help determine the extent to which outcomes have been attained for each specific program, but empirical data could not determine which outcome was

most desirable. *That is a values question.* An evaluation facilitator can help users clarify their value premises, but because the three teacher-center models were different, evaluation criteria for effectiveness varied for each type. In effect, three quite different theories of teacher development were operating in quite different educational environments. Attention to divergent theories of action helped avoid inappropriate comparisons and reframed the evaluation question from *Which model is best?* to *What are the strengths and weaknesses of each approach, and which approach is most effective for what kinds of educational environments?* Very different evaluation questions!

Matching a Theory of Change with Levels of Evidence

Claude Bennett (1982, 1979) conceptualized a relationship between the "chain of events" in a program and the "levels of evidence" needed for evaluation that became widely used and, in his honor, is known as "Bennett's Hierarchy." Although his work was originally aimed at evaluation of cooperative extension programs (agriculture, home economics, and 4-H/youth programs), his ideas are generally applicable to any education-oriented intervention. Exhibit 10.6 depicts Bennett's model.

The model presents a typical chain of program events.

1. Inputs (resources) must be assembled to get the program started.
2. Activities are undertaken with available resources.
3. Program participants (clients, students, beneficiaries) engage in program activities.
4. Participants react to what they experience.
5. As a result of what they experience, changes in knowledge, attitudes, and skills occur (if the program is effective).

EXHIBIT 10.5

Variations in Types of Teacher Centers

Type of Center	*Primary Processes For Working with Teachers*	*Primary Outcomes of the Process*
1. Behavioral centers	Curriculum specialists directly and formally instruct administrators and teachers.	Adoption of standardized curriculum systems, methods, and packages by teachers.
2. Humanistic centers	Informal, nondirected teacher exploration; "teachers select their own treatment."	Teachers feel supported and important; pick up concrete and practical ideas and materials for immediate use in their classroom.
3. Developmental centers	Advisers establish warm, interpersonal, and directive relationship with teachers working with them over time.	Teachers' thinking about what they do and why they do it is changed over time; individualized teacher personal development.

SOURCE: Adapted from Feiman (1977).

Theory-Practice Connection

Nothing as practical as a good theory.

Carol Weiss (1995:1)

It is sometimes said that there are two kinds of people in the world: thinkers and doers. And, of course, the third type: those who neither think nor do, but we won't worry about them just now. Thinkers are the world's theoreticians. They love ideas, many of which have yet to be tested and may prove quite impractical. Doers, on the other hand, are too busy doing to worry about theory. But ultimately, theory and practice ought to connect. Practice is the test of theory. Theory is the explanation of practice.

The evaluator's job is to challenge both practitioners and theoreticians. With the latter we ask, "So, it works in theory, but does it work in practice?" And with practitioners we ask, smiling diabolically, "Yes, it works in practice, but does it work in theory?"

6. Behavior and practice changes follow knowledge and attitude change.
7. Overall community impacts result as individual changes accumulate and aggregate—both intended and unintended impacts.

Bennett's Hierarchy is a values hierarchy because the model explicitly places higher value on higher-level results. The hierarchy places the highest value on attaining ultimate social and economic goals (e.g., increased agricultural production, increased health, or a higher quality of community life). Actual adoption of recommended practices and specific changes in client behaviors are necessary to achieve ultimate goals and are valued over knowledge, attitude, and skill changes. People may learn about some new agricultural technique (knowledge change), believe it's a good idea (attitude change), and know how to apply it (skill change)—but the higher-level criterion is whether they actually begin using the new technique (i.e., change their agricultural practices). Participant reactions (satisfaction, likes, and dislikes) are lower on the hierarchy. All these are outcomes, but they are not equally valued outcomes. The bottom part of the hierarchy identifies the means necessary for accomplishing higher-level ends; namely, in descending order, (3) getting people to participate, (2) providing program activities, and (1) organizing basic resources and inputs to get started.

Evaluation Use Theory of Change

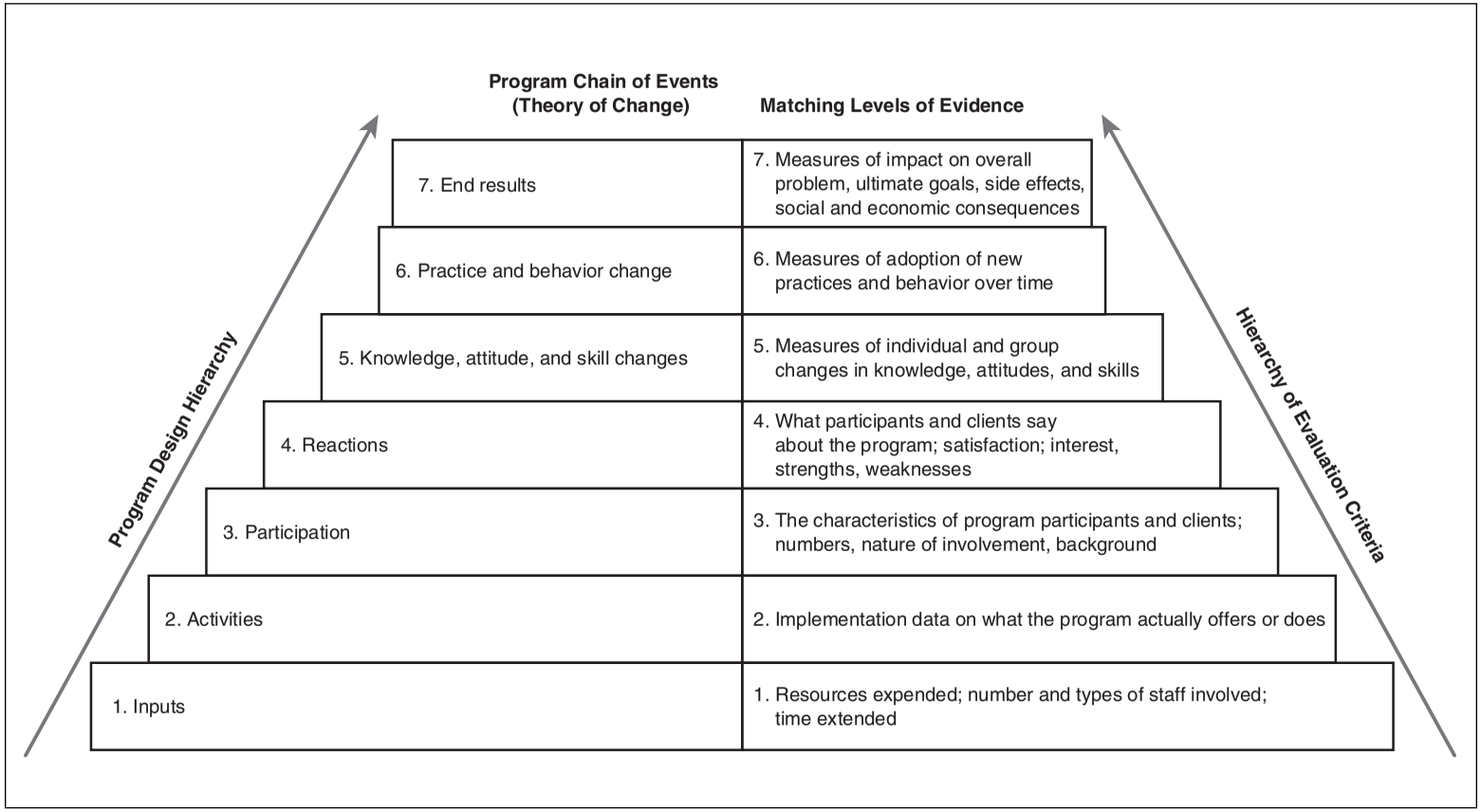
Interestingly, this same hierarchy can be applied to evaluating evaluations. Exhibit 10.7 shows a hierarchy of evaluation accountability. In utilization-focused evaluation, the ultimate purpose of evaluation is to improve programs and increase the quality of decisions made.

To accomplish this ultimate end, a chain of events must unfold.

1. Resources must be devoted to the evaluation, including stakeholder time and financial inputs.
2. Working with intended users, important evaluation issues are identified and questions focused; based on those issues and questions, the evaluation is designed and data are collected.

EXHIBIT 10.6

Theory-Evidence Hierarchy



3. Key stakeholders and primary users are involved throughout the process.
4. Intended users react to their involvement (ideally in positive ways).
5. The evaluation process and findings provide knowledge and new understandings.
6. Intended users interpret results, generate and adopt recommendations, and use evaluation results.
7. The program improves, and wise decisions are made.

Each step in this chain can be evaluated. Exhibit 10.7 shows the evaluation question that corresponds to each level in the evaluation use logic model.

Cautions against Theory of Change Work

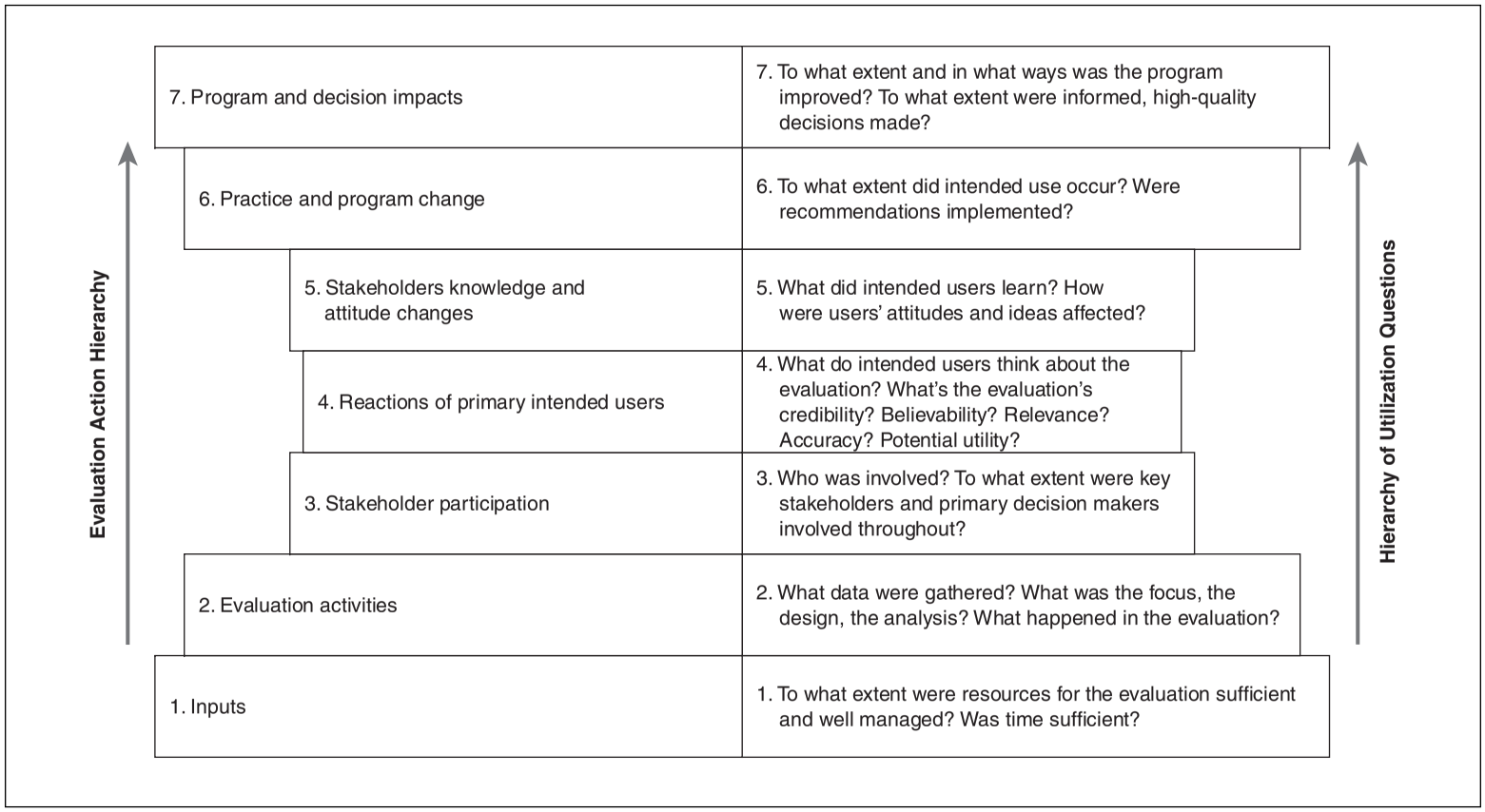
Eminent evaluation theorist Michael Scriven warns evaluators against thinking that logic modeling and theory testing are central to the work of evaluation. He wants evaluators to stay focused on the job of judging a program's merit or worth. From his perspective, rather than being essential, elucidating program theory is "a luxury for the evaluator." He considers it "a gross though frequent blunder to suppose that one needs a theory of learning to evaluate teaching" (1991b:360). One does not need to know anything at all about electronics, he observes, to evaluate computers. He also cautions that considerable time and expense can be involved in doing a good job of developing and testing a program theory. Because program theory development is really program development work rather than evaluation work, he would prefer to separate the cost of such work from the evaluation budget and scope of work. In commenting on the American Evaluation Association (AEA) listserv, *Evaltalk,* Scriven, with his renowned wit, offered this advice to evaluators:

> First, there's no law or ethical reason why evaluators with the appropriate training can't do other things for their clients besides evaluation, e.g., market surveys or statistical analyses or logic models. The law and ethics only require that they not argue that these are an essential part of evaluation. . . . There are many cases where it is beyond the boundaries of scientific knowledge to provide the logic model, and in many others it's a huge task that risks jeopardizing the primary task of evaluation.
>
> Second, this isn't ALWAYS so, and when it can be done without undue cost, helping the program managers improve their logic models—which THEY certainly need—is a Good Thing to do. Prudence requires you to keep in mind that doing this will, on some occasions (if it is possible at all), completely antagonize the clients, since they are wedded to a type of logic model that evidence does not support; prayer is one such case, but there are many, many more where the model is essentially "their thing," into which their ego is woven. So this unnecessary extra task CAN cost you the contract, or the rehire; think twice before doing it IN PUBLIC. (Scriven 2006a)

Theory-driven evaluations can also seduce evaluators away from answering straightforward formative questions or making summative judgments into the ethereal world of academic theorizing. While theory construction is a mechanism by which evaluators can link program evaluation findings to larger social scientific issues for the purpose of contributing to scientific knowledge, when conducting a utilization-focused evaluation the initial theoretical formulations originate with primary stakeholders and intended users; scholarly interests are adapted to the evaluation

EXHIBIT 10.7

Evaluating Evaluation: Logic Model of Use



needs of relevant decision makers, not vice versa. Attention to theoretical issues can provide useful information to stakeholders when *their* theories are formulated and reality-tested through the evaluation process. As always, the decision about whether and how much to focus the evaluation on testing the program's theory is driven by the question of utility: Will helping primary intended users elucidate and test their theory of change lead to program improvements and better decisions. And as Scriven adjures, ask the cost-benefit question: Will the program theory work yield sufficient benefits to justify the likely added costs involved in such work?

As advocates of theory-driven evaluation assert, a better understood program theory can be the key that unlocks the door to effective action. But how much to engage stakeholders and intended users in articulating their theories of change is a matter for negotiation. Helping practitioners test their espoused theories and discover real theories-in-use can be a powerful learning experience, both individually and organizationally. The delineation of assumed causal relationships in a chain of hierarchical objectives can be a useful exercise in the process of focusing an evaluation. It is not appropriate to construct a detailed program theory for every evaluation situation, but it is important to consider the option. Therefore, the skills of a utilization-focused evaluation facilitator include being able to help intended users construct a means-ends hierarchy, specify validity assumptions, link means to ends, and lay out the temporal sequence of a hierarchy of objectives.

But that's not all. In the last decade, the options for conceptualizing and mapping program theories have expanded and now include bringing systems perspectives into evaluation. The remainder of this chapter presents, illustrates, and discusses what it means to bring systems thinking to bear in evaluating theories of change.

Systems Theory and Evaluation

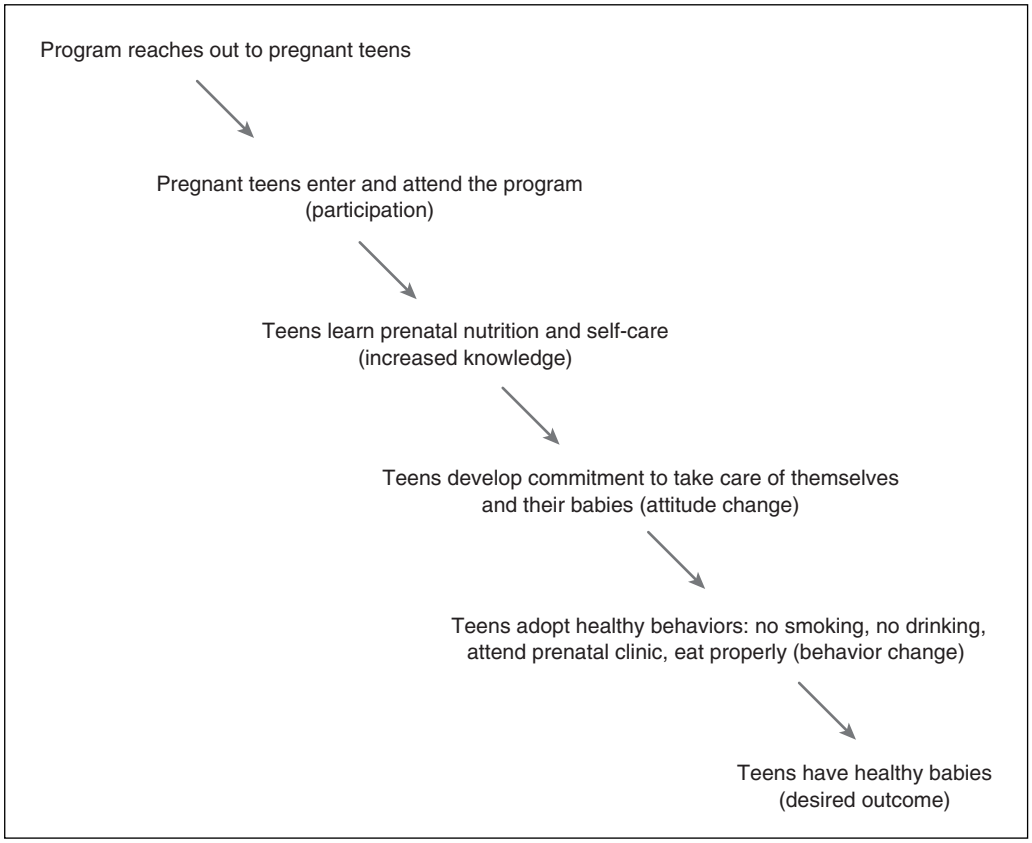
> *All models are wrong, but some are useful.*
>
> —George Box
> (quoted by Berk 2007:204)

Let's look at how modeling a program using systems thinking changes a program theory. We'll use as an example a program for pregnant teenagers. The purpose of the program is to teach pregnant teenagers how to take care of themselves so that they have healthy babies. Exhibit 10.8 shows a classic linear logic model for such a program. The teenager learns proper prenatal nutrition and self-care (increased knowledge), which increases the teenager's commitment to taking care of herself and her baby (attitude change), which leads to changed behavior (no smoking drinking, or drug use; eating properly and attending the prenatal clinic regularly). This is a linear model because *a* leads to *b* leads to *c*, et cetera: Program participation leads to knowledge change, which leads to attitude change, which leads to behavior change, which produces the desired outcome (a healthy baby). This is a linear cause-effect sequence as depicted in Exhibit 10.8. This is the traditional, widespread approach to logic modeling.

Now, let's ask some systems questions. What various influences actually affect a pregnant teenager's attitudes and behaviors? The narrowly focused, linear model in Exhibit 10.8 focuses entirely on the program's effects and ignores the rest of the teenager's world. When we ask about that world, we are inquiring into the multitude of relationships and connections that may

EXHIBIT 10.8

Linear Program Logic Model for Teenage Pregnancy Program



influence what the pregnant teenager does. Exhibit 10.9 is a rough sketch of possible system connections and influences. We know, for example, that teenagers are heavily influenced by their peer group. The linear, narrowly focused logic model, targets the individual teenager. A systems perspective that considered the influence of a pregnant teenager's peer group might ask how to influence the knowledge, attitudes, and behaviors of the entire peer group. This would involve changing the subsystem (the peer group) of which the individual pregnant teenager is a part. Likewise, the system's web of potential influences in Exhibit 10.9 invites us to ask about the relative influence of the teenager's parents and other family members, the pregnant teenager's boyfriend (the child's father), or teachers and other adults, as well as the relationship to the staff of the prenatal program. In effect, this systems perspective reminds us that the

behavior of the pregnant teenager and the health of her baby will be affected by a number of relationships and not just participation in the prenatal program. In working with such a model with program staff, the conceptual elaboration of the theory of change includes specifying which direction arrows run (one way or both ways, showing mutual influence), which influences are strong (heavy solid lines) versus weak (dotted lines), and which influences are more dominant (larger circles versus smaller circles).

Exhibit 10.10 changes the conceptualization of the change process from a simple linear model of cause-effect to a systems dynamics model of reinforcing actions that depicts the change process as cumulative. When the pregnant teenager adopts healthy behaviors, she feels better; if and when she gets positive reinforcement from clinics nurses, family, and friends, her healthy behaviors are affirmed and reinforced, and she is more likely to continue them. The causal mechanism for sustained change in Exhibit 10.10 is ongoing positive

EXHIBIT 10.9

Systems Web Showing Possible Influence Linkages to a Pregnant Teenager

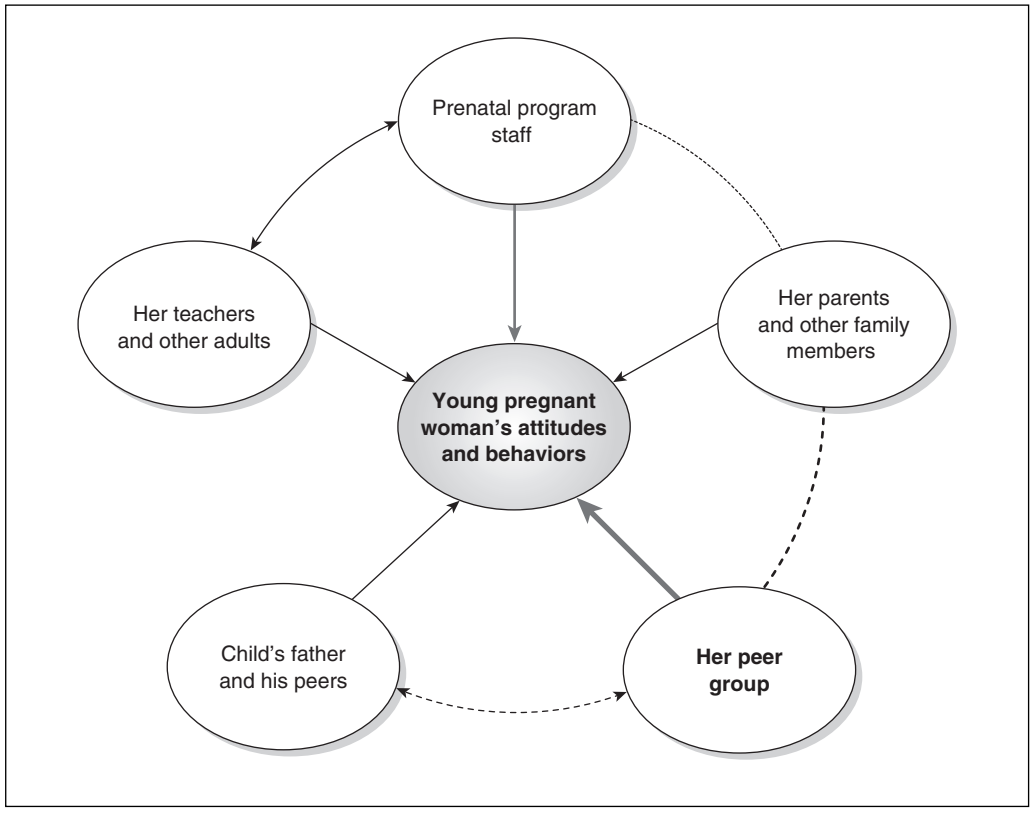
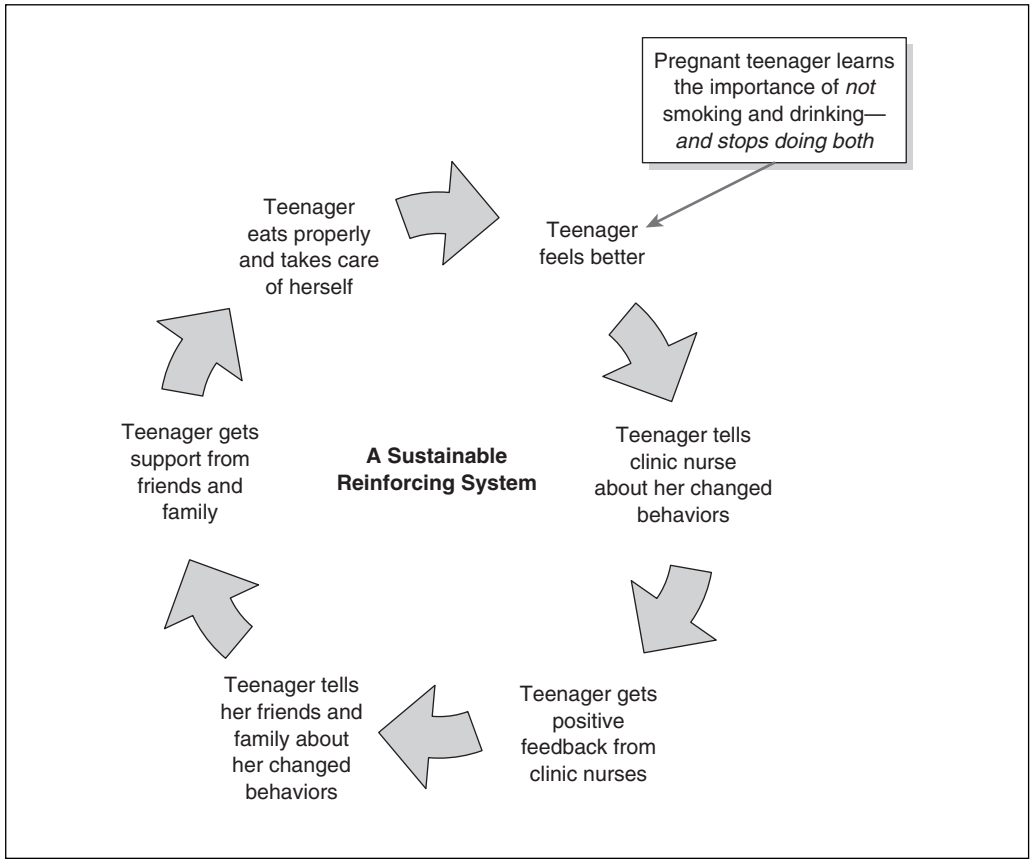


EXHIBIT 10.10

Sustainable Change: Systems Dynamic Reinforcing Feedback Loops



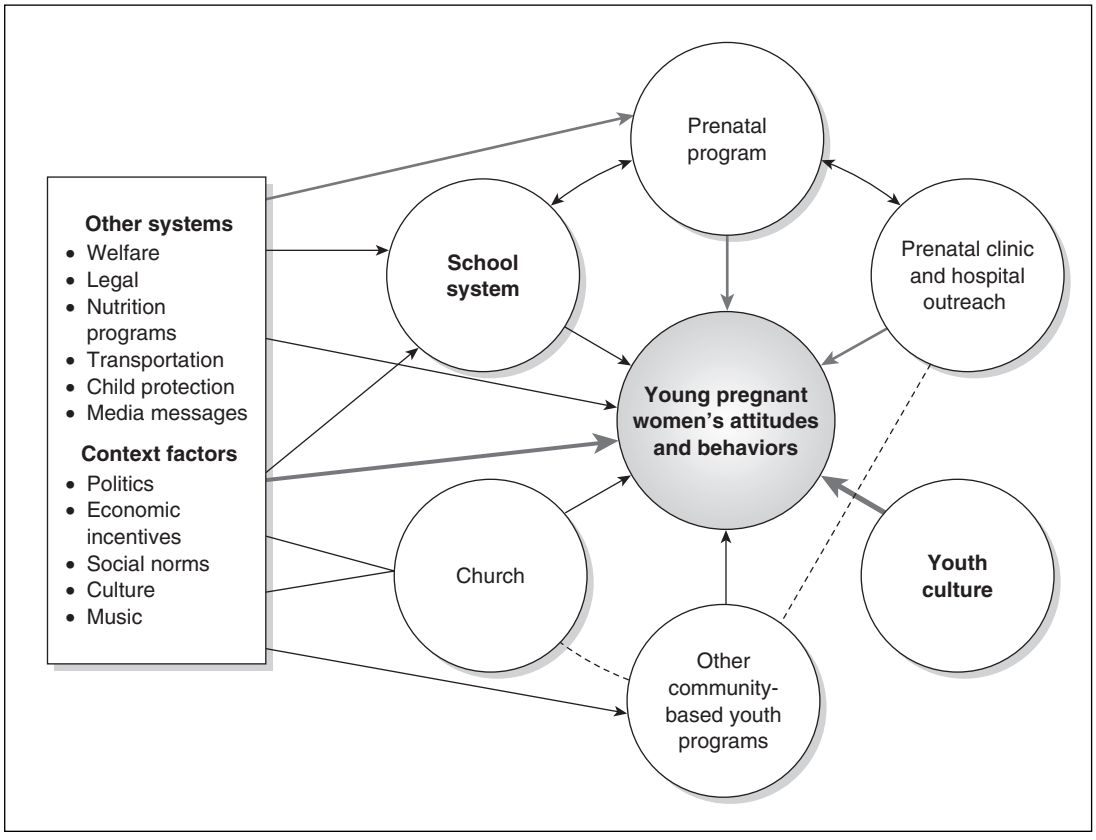
reinforcement from people who are important to the pregnant teenager, which is a different program theory from the linear model of knowledge alone leading to behavior change. Exhibit 10.10 guides the evaluator to ask not just how to produce the desire behavior—but how to sustain it. *How change is sustained is a systems question.*

Exhibit 10.11 presents yet another systems perspective depicting possible institutional influences affecting pregnant teenagers' attitudes and behaviors. The narrowly focused, linear logic model in Exhibit 10.8 treats the program's impact in isolation from other institutional and societal factors. In contrast, the systems web in Exhibit 10.11 shows the prenatal program as one potentially strong influence on pregnant teenagers but also takes into account the important influences of the youth

culture, the school system, other community-based youth programs, the local clinic and hospital, and possibly the local church. Moreover, during her pregnancy the teenager may be affected by other systems: the welfare system (eligibility for financial support and food stamps), the legal system (laws governing the degree to which the teenager can make independent decisions or live on her own), nutrition programs that might collaborate with the prenatal program, the transportation system (which affects how the teenager gets to clinic visits and the program), and the pervasive influences of the media (television, movies, music) that affect teenager attitudes and behaviors. The systems diagram in Exhibit 10.11 also includes larger contextual factors such as the political environment; economic incentives that can affect a teenager's ability to live independently, get child care, continue to attend school, or get a job; and social norms and larger cultural influences that affect how society responds to a teenager's pregnancy.

EXHIBIT 10.11

Program Systems Web Showing Possible Institutional Influences Affecting Pregnant Teenagers' Attitudes and Behavior



Constructing such a systems map with a prenatal program may lead the program to consider a more collaborative effort in which various institutional partners come together to work toward the desired outcome of a healthy pregnant teenager who delivers a healthy baby. The system diagrams suggest that the prenatal program by itself, focusing only on the teenager and only on its own delivery of knowledge to the teenager, is less likely to achieve the desired outcome than a model which takes into account the influences of other people in the teenager's system (the teenager's world) and collaborates with other institutions that can have an effect on the attainment of desired outcomes.

Systems Framework Premises

Looking at a program from a systems perspective is one way to deepen our understanding of the program and its outcomes. A systems framework is built on some fundamental relationships premises. We'll examine those premises using the program for pregnant teenagers to illustrate each one.

1. *The whole is greater than the sum of the parts.* If you look at Exhibit 10.9, you see interconnected parts—the pregnant teenager, her peer group, her family, her boyfriend, teachers and other adults who interact with her, and program staff. This *whole web of relationships* will be a unique constellation of interactions for each pregnant teenager. The *whole* may include consistent or contradictory messages about the teenager and her pregnancy. Moreover, that web of relationships isn't just built around her pregnancy. Messages about school, family, life, work, and love are all part of the mix. The systems picture reminds us that the teenager's life consists of a number of relationships and issues that extend well beyond her experience in the prenatal program. The program is but *one influence in her whole life.* Linear program theories tend to be conceptualized as if the program is the only thing going on in the participant's life. Looking at a program as but one part of a participant's whole life is what Saville Kushner (2000) has described as "personalizing evaluation."

> I will be arguing for evaluators approaching programs through the experience of individuals rather than through the rhetoric of program sponsors and managers. I want to emphasize what we can learn about programs from Lucy and Ann. . . . So my arguments will robustly assert the need to address "the person" in the program. (Pp. 9–10)

Exhibit 10.9 is an abstract conceptualization of a set of relationships. If we put Lucy in the circle depicting the pregnant teenager and capture her story as a case study, we get a more holistic understanding of Lucy's life and where the program fits in Lucy's life. We then do the same thing for Ann. Each of those stories is its own whole, and the combination of stories of teenagers in the program gives us a sense of the program whole. But that program whole cannot be reduced to the individual stories (the parts) any more than Lucy's life can be reduced to the set of relationships in Exhibit 10.9. The whole is greater than the sum of parts. Moreover, Exhibit 10.11 reminds us that the program cannot be understood as a free-standing, isolated entity. The program as a whole includes relationships with other entities—schools, community organizations, churches—and larger societal influences. A systems framework invites us to understand the program in relation to other programs and as part of a larger web of institutions.

2. *Parts are interdependent such that a change in one part has implications for all*

parts and their interrelationships. Imagine that when Lucy becomes pregnant and enters the prenatal program, she has a close relationship with her boyfriend (the child's father) and her family, is doing well in school, and is active in church. The stress of the pregnancy leads Lucy and her boyfriend to break up. Things become tense in her family as everyone wants to give her advice. Her school attendance becomes irregular and she stops going to church. Without her boyfriend and with increased family tension, a small number of female peers become increasingly central to Lucy's life. What we begin to understand is that Lucy's *system of relationships* existed before she became pregnant. Her pregnancy affects all her relationships and those changes in relationships affect each other, ebbing and flowing. The pregnancy can't really be said to "cause" these changes. What happens between Lucy and her boyfriend when she becomes pregnant is a function of their whole relationship and their relationships with others. The program is part of that mix—but only a part. And how Lucy experiences the program will be affected by the other relationships in her life.

3. *The focus is on interconnected relationships*. The change in perspective that comes with systems thinking focuses our attention on how the web of relationships function together rather than as a linear chain of causes and effects. It is different to ask how things are connected than to ask does *a* cause *b*. It's not that one inquiry is right and the other is wrong. The point is that different questions and different frameworks provide different insights. Consider the example of your reading this book. We can ask, "To what extent does reading this book increase your knowledge of evaluation?" That's a fairly straightforward linear evaluation question. Now we ask, "How does reading this book *relate to* the other things going on in your life?" That's a simple systems question. Each question has value, but the answers tell us very different things.

4. *Systems are made up of subsystems and function within larger systems*. Exhibit 10.9 shows a pregnant teenager's relationships with other people. Exhibit 10.11 shows the program's relationships with other institutions and how these, in combination, influence the teenager's attitudes and behavior. The "subsystems" are the various circles in these two exhibits. These subsystems—family, school, church, community, peer group—function within larger systems such as society, the legal system, the welfare system, culture, and the economy. How subsystems function within larger systems and how larger systems connect to and are influenced by subsystems can be part of a systems inquiry into understanding a program and its effects. Both the content and processes of a prenatal program for pregnant teenagers will be affected by larger societal norms. That's why programs in rural Mississippi, inner city Chicago, East Los Angeles, southern France, northern Brazil, and Burkina Faso would be different—*even if they supposedly were based on the same model*. The societal and cultural contexts would inevitably affect how the programs functioned.

5. *Systems boundaries are necessary and inevitably arbitrary*. Systems are social constructions (as are linear models). Systems maps are devices we construct to make sense of things. It is common in hiking to remind people in the wilderness that "the map is not the territory." The map is an abstract guide. Look around at the territory. What to include in a systems diagram and where to draw the boundaries are matters of utility. Including too much

makes the system overwhelming. Including too little risks missing important elements that affect program processes and outcomes. Given the purpose of evaluation, to inform judgment and action, the solution is to be practical. If we are mapping the relationships that affect a teenager's health during her pregnancy (Exhibit 10.9), we ask, "What are the primary relationships that will affect what the teenager does?" List those and map them in relationship to each other. Don't try to include every single relationship (a distant cousin in another city with whom she seldom interacts), but include all that are important—including that distant cousin if she is a teenager who has recently been through a pregnancy, which might be represented by a circle designating "other teenagers who have been or are pregnant who this teenager knows." The systems map of a program is a guide, a way to ask questions and understand the dynamics of what occurs. The systems map is not, however, the program.

A first step in moving beyond simple linear logic models is to add feedback loops to the model. Once a program is in operation, the relationships between links in the causal hierarchy are likely to be recursive rather than unidirectional. Instead of *a* causes *b*, the model becomes *a* causes *b*, and achieving *b* stimulates more of *a*. Take dieting and exercising to lose weight. Eating less and exercising leads to weight loss; losing weight leads to looking and feeling better. That's the simple linear chain. But looking and feeling better reinforces eating better and continuing to exercise. See Exhibit 10.12.

Systems Framework Premises

A systems framework is built on some fundamental relationships premises.

- The whole is greater than the sum of the parts.
- Parts are interdependent such that a change in one part has implications for all parts and their interrelationships.
- Systems are made up of subsystems and function within larger systems.
- The focus is on interconnected relationships among parts, and between parts and the whole.
- Systems boundaries are necessary and inevitably arbitrary.

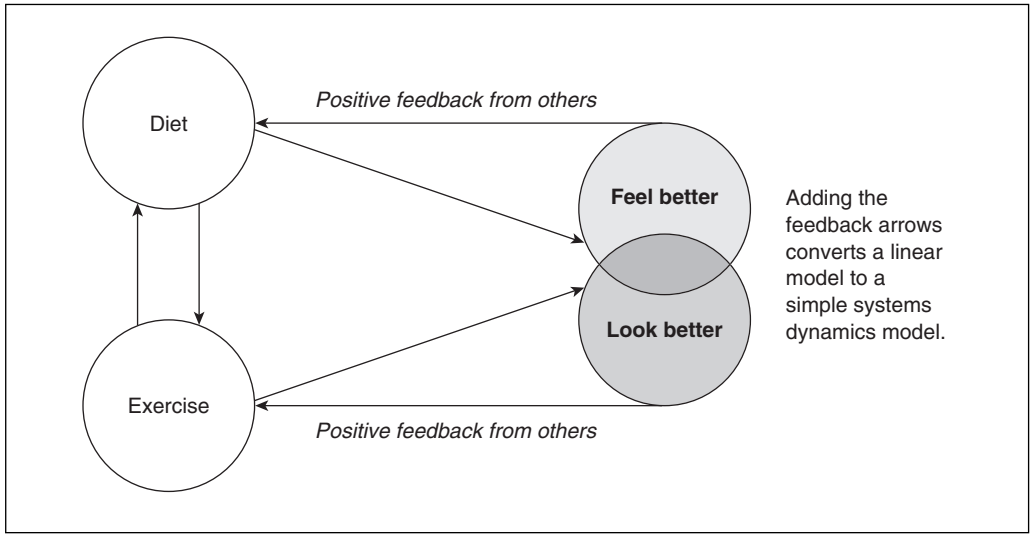
Recognizing such recursive feedback effects changes the model. For example, high-achieving schools affect the opinions and actions of parents, but parent reactions also affect the degree to which schools are committed to high achievement. The influence doesn't flow just one way. Classroom climate and school curriculum affect student achievement, but variations in student achievement also affect school climate and curriculum. From a systems perspective, a simple linear means-ends hierarchy without feedback loops or interdependencies is likely to be oversimplified, but there is no avoiding some simplification, even with systems maps. The basic dilemma is how much to simplify reality. *The challenge is to construct simplifications that pass the dual tests of usefulness and accuracy.*

The Increasing Importance of Systems Thinking in Evaluation

The preceding has provided only a brief introduction to the possibilities for

EXHIBIT 10.12

Converting a Linear Model to a Simple Feedback Model



incorporating systems perspectives in evaluation. At the 2002 national conference of the AEA, president Molly Engle made systems the focus of the meeting. The keynote address by noted systems thinker John Sterman was titled "No Learning Without Feedback: The Vital Partnership of Evaluation and Systems Thinking." Shortly thereafter a Topical Interest group focused on Systems was formed within AEA and every national conference since has had a full strand of sessions devoted to systems approaches to evaluation. In 2006, AEA published its first-ever monograph, an expert anthology on *Systems Concepts in Evaluation* edited by Bob Williams and Iraj Iman. That monograph provides a wide range of systems approaches and demonstrates the diversity of approaches that congregate under the systems umbrella. In commenting on this diversity the editors wrote,

> For those of you looking for coherence about what we consider to be relevant systems concepts for evaluation, our advice when reading this publication is to look for patterns rather than definitions. For us, three patterns stand out:
>
> 1. *Perspectives.* Using systems concepts assumes that people will benefit from looking at their world differently. For systems practitioners, this motivation is explicit, deliberate, and is fundamental to their approach. However, just looking at the "bigger picture," or exploring interconnections does not make an inquiry "systemic." What makes it systemic is how you look at the picture, big or small, and explore interconnections. A "system" is as much an

"idea" about the real world as a physical description of it.

2. *Boundaries.* Boundaries drive how we "see" systems. Boundaries define who or what lies inside and what lies outside of a particular inquiry. Boundaries delineate or identify important differences (i.e., what is "in" and what is "out"). Boundaries determine who or what will benefit from a particular inquiry and who or what might suffer. Boundaries are fundamentally about values–they are judgements about worth. Defining boundaries is an essential part of systems work/inquiry/thinking.

3. *Entangled systems.* One can observe and perceive systems within systems, systems overlapping other systems, and systems tangled up in other systems. Thus it is unwise to focus on one view or definition of a system without examining its relationship with another system. Where does one system begin and the other end? Is there overlap? Who is best situated to experience or be affected by that overlap? What systems exist within systems and where do they lead? A systems thinker always looks inside, outside, beside, and between the readily identified systems boundary. He or she then critiques and if necessary changes that initial choice of boundary. (Williams and Iman 2006:6)

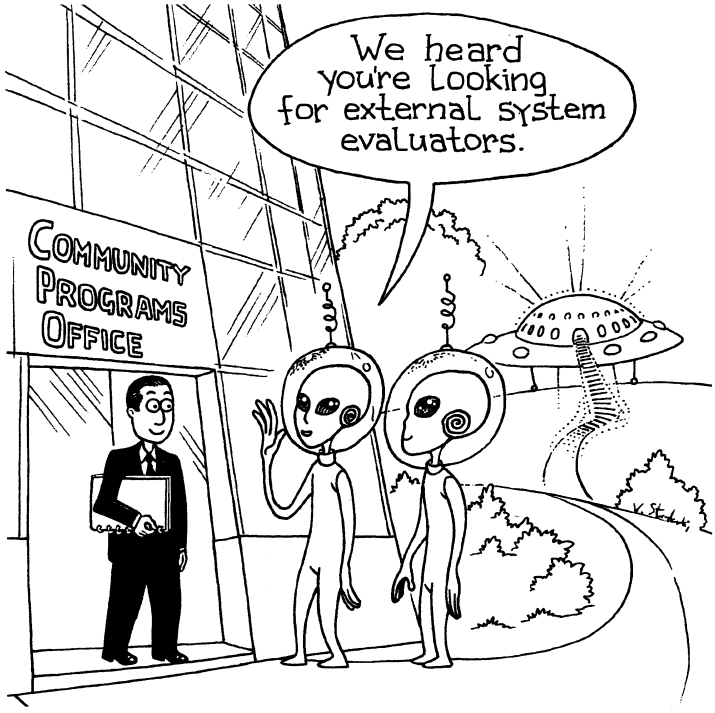
Part of the challenge of incorporating systems thinking in evaluation is that there are so many different systems meanings, models, approaches, and methods, including system dynamics, soft systems methodology, cultural-historical activity theory, and critical systemic thinking, each of which has specific implications for evaluation (Williams 2005b). More generally, critical systemic thinking can be considered as one element of *evaluative thinking* (see Chapter 5). Werner Ulrich (2000, 1998) has advocated that although not everyone should become a systems scholar, systems thinking involves new reflective skills that are essential in modern society for both professional competence and effective citizenship.

Evaluating Systems Reform

Systems thinking is pushing evaluators to conceptualize what we do in new ways and offers new frameworks for use in working with primary intended users to think about what they do and how they do it. This is especially the case where the targeted unit of change is, itself, *a system.* Thus, while much program evaluation has traditionally focused on the outcomes of programs aimed at individuals—students, farmers, chemically dependent people, parents, children, professionals, some initiatives target systems for reform and change. Policy initiatives can be aimed at reforming systems: the health care system, the educational system, the judicial system, the farming system, et cetera. Evaluating advocacy initiatives and policy change campaigns changes the unit of analysis (the *evaluand*) from the program level to the policy or systems level (Patton 2008, Coffman 2007a, b). While systems thinking is an option in looking at program outcomes for individuals, it is essential for evaluating system reform initiatives. And that provides a segue to one particularly challenging systems evaluation problem: How to evaluate emergent processes in complex nonlinear systems.

Evaluation in Complex Adaptive Systems

A Complex Adaptive System is a dynamic network of many interacting parts, continuously acting and reacting. The results of these interactions are dynamic, emergent, uncertain, and unpredictable. Examples are weather systems, stock markets, ecosystems, and anthills. One of the characteristics of complex adaptive systems is that small effects can have large consequences as



expressed by the butterfly effect metaphor, which suggests that a butterfly flapping its wings today in China may lead to a typhoon forming in the Pacific Ocean months later. This is represented in our everyday experience by the story of a chance, brief encounter that changes your life, or a phrase offered at an opportune moment that turns you in a new direction and alters forever your path. Since the 2000 presidential election, the butterfly effect has an additional meaning. Election officials in Palm Beach, Florida, experimented with a new format and procedure for voting called the "butterfly ballot." This resulted in voter confusion in a tight election, propelled the election results into the judicial system where the Supreme Court refused to allow a recount giving Florida to George Bush, which gave the presidency to George Bush, which led to the Iraq invasion, which led to global events and processes still unfolding. Small actions (a changed ballot in one county) can have huge repercussions as that action reverberates through a complex adaptive system.

Sometimes complex effects take years. In 1933, the Belgians who controlled Rwanda as a colony issued identity cards classifying every Rwandan as Tutsi, Hutu, or Twa (a very minor category). In 1994, as part of the tragic genocide in Rwanda, those cards were used by Hutu to identify hundreds of thousands of Tutsi and kill them (Kinzer 2007:24). How does one portray such a connection? Certainly, not simple cause and effect. And somehow more than merely unintended consequences. A complex nonlinear system unfolded in an unpredictable fashion.

Complexity science is being used to understand phenomena in the biological world, policy analysis, ecosystems, economic systems, and in organizations (Fritjof et al. 2007; Dennard, Richardson, and Morcol 2005;

Richardson et al. 2005; Gribben 2004; Westley and Miller 2003; Gunderson and Holling 2002; Johnson 2001; Lewin 2001; Zimmerman, Lindbery, and Plsek 2001; Eoyang 1996; Waldrop 1992). *But what does this have to do with evaluation?* The answer lies in situational responsiveness and problem definition, which affect how we conceptualize and design evaluations.

Three Kinds of Problems: Simple, Complicated, Complex

> *To pursue greatness is to pursue* Maybe.
>
> —John Bare (2007), Vice President
> The Arthur M. Blank
> Family Foundation
> Atlanta, Georgia

In studying social innovations, we were impressed by the uncertainty and unpredictability of the innovative process, even looking back from a mountaintop of success, which is why we called the book *Getting to Maybe* (Westley, Zimmerman, and Patton 2006). Evaluating social innovations is a complex problem, as opposed to evaluating simple and complicated problems. A *simple* problem is how to bake a cake following a recipe. A recipe has clear cause-and-effect relationships and can be mastered through repetition and developing basic skills. There is a chance to standardize the process and to write the recipe with sufficient detail that even someone who has never baked has a high probability of success. Best practices for programs are like recipes in that they provide clear and high fidelity directions since the processes that have worked to produce desired outcomes in the past are highly likely to work again in the future. Assembly lines in factories have a "recipe" quality as do standardized school curricula. Part of the attraction of the 12-Step program of Alcoholics Anonymous is its simple formulation.

A *complicated* problem is more like sending a rocket to the moon. Expertise is needed. Specialists are required and coordination of the experts is another area of expertise itself. Formulae and the latest scientific evidence are used to predict the trajectory and path of the rocket. Calculations are required to ensure sufficient fuel based on current conditions. If all the "homework" is completed, and if the coordination and communication systems are sophisticated enough to access the expertise, there is a high degree of certainty of the outcome. It is *complicated*, with many separate parts that need coordination, but it can be controlled by knowledgeable leaders and there is a high degree of predictability about the outcomes. Cause-and-effect relationships are still very clear, although not as straightforward as with simple problems. Coordinating large-scale programs with many local sites throughout a country or region is a complicated problem.

Parenting is *complex*. Unlike the recipe and rocket examples, there are no clear books or rules to follow to guarantee success. Clearly, there are many experts in parenting and many expert books available to parents. But none can be treated like a cookbook for a cake, or a set of formulae to send a rocket to the moon. In the case of the cake and the rocket, for the most part, we were intervening with inanimate objects. The flour does not suddenly decide to change its mind, and gravity can be counted on to be consistent too. On the other hand, children, as we all know, have minds of their own. Hence our interventions are always in relationship with them. There are very few stand-alone parenting tasks. Almost always, the parents and child interact to create outcomes. *Any highly*

individualized program has elements of complexity. The outcomes will vary for different participants based on their differing needs, experiences, situations, and desires.

Exhibit 10.13 highlights the distinctions between the three kinds of problems. In all three cases, we tend to be optimistic that positive outcomes can be achieved. However, the way we intervene in each of these contexts is qualitatively different, as is how we design an evaluation (Rogers 2008; Westley et al. 2006:8–10).

Simple formulations invite linear logic models that link inputs to activities to outputs to outcomes like a formula or recipe. Complicated situations invite system diagrams and maps that depict the relationships among the parts. Complex problems and situations are especially appropriate for *developmental evaluation* in which the

EXHIBIT 10.13

Simple, Complicated, and Complex Lenses

Simple	*Complicated*	*Complex*
Following a recipe	*Sending a rocket to the moon*	*Raising a child*
The recipe is essential.	Right protocols or formulae are critical and necessary.	Rigid protocols have a limited application or are counter-productive.
Recipes are tested to assure easy replication.	Sending one rocket to the moon increases assurance that the next will be also be a success.	Raising one child provides experience but is no guarantee of success with the next.
No particular expertise is required but cooking expertise increases success rate.	High levels of expertise and training in a variety of fields are necessary for success.	Expertise helps but only when balanced with responsiveness to the particular child.
A good recipe produces nearly the same cake every time.	Key elements of each rocket MUST be the same to succeed.	Every child is unique and must be understood as an individual.
The best recipes give good results every time.	There is a high degree of certainty of outcome.	Uncertainty of outcome remains.
A good recipe specifies the quantity and nature of the "parts" needed and the order in which to combine them, but there is room for experimentation.	Success depends on a blueprint that directs both the development of separate parts and specifies the exact relationship in which to assemble them.	Can't separate the parts from the whole; essence exists in the relationship between different people, different experiences, different moments in time.

SOURCE: Westley et al. (2006:9).

evaluation design is flexible, emergent, and dynamic, mirroring the emergent, dynamic, and uncertain nature of the intervention or innovation being evaluated. Tracking, monitoring, and evaluating complex policies requires continuous "streams of knowledge" rather than discrete and bounded studies (Stame 2006a, b). Evaluating in the face of the uncertainties of complexity requires anticipation and agility to improve evaluation quality, responsiveness, and real-time relevance (Morrell forthcoming, 2005).

Complexity scientist Ralph Stacey (2007, 1996, 1992) has offered a matrix of two dimensions that helps distinguish simple, complicated, and complex situations. One dimension scales the degree of certainty in the cause-effect relationship. Programs and interventions are close to certainty when cause and effect linkages in the logic model are highly predictable, as in the relationship between immunization and preventing disease. At the other end of the certainty continuum are innovative programs where the outcomes are highly unpredictable; a community development initiative would typically involve considerable uncertainty. Extrapolating from past experience is problematic because, like rearing a child, each community is unique. The vertical axis of the matrix captures the degree of agreement among various stakeholders about a program's needed inputs, goals, processes, outcomes measures, and likely long-term impacts. High levels of agreement make situations fairly simple; high degrees of values conflict foment complexity. Exhibit 10.14 shows where on this matrix the zones defining simple, complicated, and complex problems can be expected.

- Simple interventions are defined by high agreement and high causal certainty; immunization to prevent disease fits this zone on the matrix.
- Socially complicated situations are defined by fairly high predictability of outcomes, but great values conflict among stakeholders; abortion is an example.
- Technically complicated situations are defined by high agreement among stakeholders but low causal certainty; everyone wants children to learn to read but there are ferocious disagreements about which reading approach produces the best result (Schemo 2007).
- Complex situations are characterized by high values conflict and high uncertainty; what to do about global warming would fall in the complexity zone of the matrix.

Let me now explain and illustrate the evaluation implications of these different ways of understanding a program or intervention.

An Evaluation Example Illustrating Simple, Complicated, and Complex Designs

Consider a nationwide leadership development program that aims to infuse energy and vitality into a moribund nonprofit sector (a judgment based on funder assessment). The intensive 18-month program includes

(1) *skill development* (e.g., communications training, conflict resolution, needs assessment, strategic planning, appreciative inquiry methods) and knowledge acquisition (e.g., introduction to various theories of change, systems thinking, complexity science),

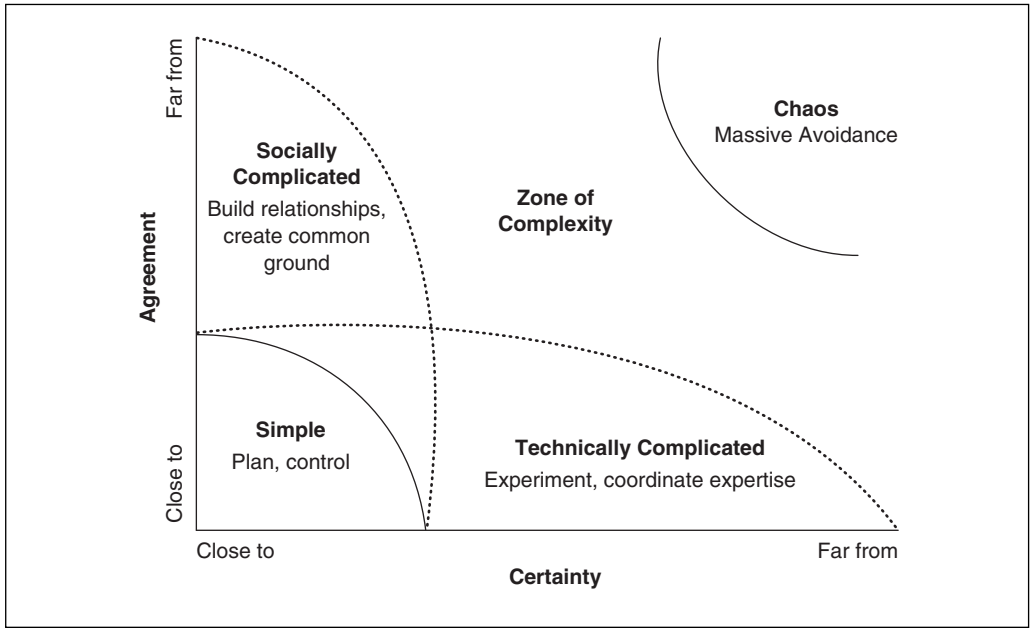
(2) *an organizational change project* in participants' own organizations, and

(3) *networking* with other participants around nonprofit sector issues of common interest and concern.

Skill development and knowledge acquisition can be modeled and evaluated

EXHIBIT 10.14

Matrix Depicting Simple to Complex



SOURCE: Adapted from Stacey (2007); Zimmerman, Lindbery and Plsek (2001:136–141).

with a linear framework. The desired outcomes are specifiable, concrete, and measurable, and the outcomes are connected directly to the curriculum and training in a short, observable time frame. Participants demonstrate their skills and knowledge by writing papers, carrying out assignments, and doing team projects. A linear logic model can appropriately capture and depict the hypothesized connections between inputs, activities, outputs, and outcomes as a framework for evaluation.

The second program component—carrying out organizational change projects in their own organizations—is congruent with relationship-focused systems modeling and systems change evaluation. The operations, culture, and constellation of units within each participant's organization constitute a baseline organizational system at the time each participant enters the leadership development program. Each organization functions within some context and environment. As part of the leadership development experience, participants undertake some self-selected change effort, e.g., board development, strategic planning, staff development, reorganization, or evaluation, among many possibilities. These are efforts aimed at increasing the effectiveness of the participating organizations and can be modeled and evaluated as systems change initiatives. Evaluative case studies would capture the changed relationships within the organizations, both changed relationships among internal elements (e.g., between board and

staff, or between technology support units and line management units) as well as changed relationships with organizations in the environment (e.g., collaborations, new or changed partnerships, new suppliers, changed community or funder relationships). The focus on changed relationships, linkages, and connections makes systems change evaluation an especially appropriate framework for this aspect of the program.

The third overall thrust of the program involves supporting self-organizing networks among participants to infuse new energies and synergies into the nonprofit sector. This constitutes a vision rather than a measurable goal. It's not at all clear what may emerge from such networking (no clear causal model), and the value of such networking is hard to measure. Indeed, there's no particular plan to support such networking other than bringing these leaders together and have them interact for intense periods of time. Beyond the networking, it's both impossible to predetermine what might occur as a result of the infusion of new leadership into the nonprofit sector and it would be politically inappropriate for the philanthropic funder to make such a determination because it would be controversial. Indeed, part of the intervention is support for the nonprofit and voluntary sector leaders to engage in dialogue around what actions and initiatives would revitalize the sector. The outcomes in this case will be entirely emergent. The evaluation would involve real-time monitoring of emergent initiatives watching for what the self-organizing networking yields. Indeed, in a real case where this form of emergent evaluation was actually undertaken, the results turned up conferences organized, regional institutes established, lobbying efforts coordinated, collaborations created, new partnerships, and shared development of materials. *None of these efforts were predictable in advance.* They *emerged* from the process and were captured through developmental evaluation, specifically, periodically e-mailing participants to inquire about how they were working with others and, when something turned up, interviewing them about the details. Exhibit 10.15 summarizes and compares how these three evaluation approaches, representing different theories of change, can be combined in a single comprehensive evaluation of the leadership development program.

Matching the Evaluation Framework to the Nature of the Intervention

The principle illustrated by the preceding leadership development program is that the modeling framework and evaluation approach should be congruent with the nature of a program intervention. Understanding an intervention as simple, complicated, or complex can significantly affect how an evaluation is conducted (Rogers 2008; Martin and Sturmberg 2005).

When the intervention is readily understood as fitting traditional linear logic modeling, then the evaluation would document the program's inputs, processes, outputs, outcomes, and impacts, including viable and documentable linkages connecting the elements of the model. This is the traditional and dominant approach to program theory modeling and evaluation. In most cases, the outcomes for a linear logic model will be changes at the individual level among intended beneficiaries, e.g., changes in attitudes, knowledge, behavior, and status (well-being, health, employment, etc.). In other words, the unit of analysis for the evaluation is typically individuals and individual-level change.

Systems mapping offers an approach for evaluating systems change efforts. In many cases, the object of an intervention is

EXHIBIT 10.15

Evaluation Design for a Leadership Development Program: Different Components Manifest Different Theories of Change

Program component	*Leadership development*: Increased knowledge and skills, and use of those skills in their work.	*Organizational change*: Each participant carries out a project of his or her own choosing to develop the organization.	*Networking and leadership within the national nonprofit sector*: Vision of new energy and vitality.
Problem framing	Simple/Complex	Complicated/Complex	Complex
Type of theory of change	*Linear logic model*: Program training increases knowledge and skills. *Complex*: Additional unique, unanticipated, and emergent outcomes likely with such a high-powered group.	*Systems change*: Participants' organizational systems are changed through projects. Projects vary greatly and are chosen by participants and their organizations. *Complex*: Each is unique.	*Complex adaptive self-organizing network:* Informal groups emerge and decide to collaborate around shared interests.
Degree of certainty about how to achieve desired outcomes. (Horizontal axis on the Stacey Matrix)	*High certainty*: many leadership programs have produced changes in skills & knowledge. There is substantial knowledge about how to support professional development. Highly experienced instructors.	*Moderate to low certainty*: Degree to which organizations change dependent on a large number of factors, many of which are outside the leaders' control.	*Very low certainty*: Outcomes are unclear and unspecified; not even possible to specify all the variables that come into play; high likelihood that chance encounters will play a part.
Degree of agreement about the desired outcomes (Vertical axis on the Stacey Matrix)	High agreement that nonprofit leaders should have leadership skills and professional development opportunities.	Varied agreement about the need for organizational change among participants' organizations; some are open, some resistant; most uncertain about what is involved.	Low agreement about how these leaders should engage together in the larger nonprofit sector; vague vision of engagement, but the specifics will be emergent and opportunistic.
Evaluation questions	Are the desired outcomes achieved? Can these outcomes be attributed to the program? Do the trained leaders use their new skills in their work?	What projects do participants do in their organizations? How are their organizations changed? How are relationships altered? How are the organizations' relationships with external institutions affected?	What informal groups of participants self-organize? What do these emergent subgroups do together? What impacts flow from their emergent activities? What developments occur over time?
Evaluation design	Pre-post assessment of changed knowledge and skills. Follow-up to assess application of new skills.	Case studies of participants' projects focusing on how their organizations are changed.	Developmental evaluation follow-ups to track what emerges and develops over time.

a change in a system, for example, developing an organization (an organizational system change effort) or creating collaborative relationships among organizations, or connecting organizations and communities in some new ways. The unit of analysis is the system, and the focus is on changed relationships and interconnections, which are the defining elements of how the system functions. For example, philanthropic funders or government grants often seek to create genuinely collaborative relationships among organizations operating independently of each other (often referred to at the baseline as operating in "silos" or "elevator shafts"). An evaluation looking at the effectiveness and results of such a systems change initiative would map changes in relationships. Network analysis and mapping (McCarty et al. 2007; Durland and Fredericks 2005; Bryson et al. 2004) are powerful tools for capturing and depicting such dynamic and evolving system relationships (or absence of same when the change process fails to work).

Outcome mapping (International Development Research Centre 2007) is an approach that recognizes the complex nature of international development initiatives, especially those that involve multiple partners collaborating together, where each is contributing in some way to changes in the behaviors, relationships, activities, or actions of the people, groups, and organizations with whom a program works directly. *Outcome mapping* incorporates aspects of linear logic modeling by having a program map the behavior changes it can affect within its direct sphere of influence; it incorporates systems thinking by mapping relationships with partners and recognizing the impossibility of drawing simple causal attribution conclusions where many factors affect change and cooperating partners each make *contributions* to change; and it brings an appreciation of complex adaptive processes to bear by treating an outcome map as emergent and dynamic, a map that can be and should be revisited and revised as conditions change (as they surely will) and as new understandings emerge (which is surely to be hoped for and encouraged).

This last point deserves emphasis. Logic models and program theories can change and evolve over the life of a program. Nick Tilley (2004) has described how the Crime Reduction Programme in the United Kingdom metamorphosed over its 3-year life, including changing theories of the program and changing forms of evaluation. The more volatile the environment of a program, the more likely it is that the program theory will be affected by that volatility—and by learning from what actually unfolds.

Developmental Evaluation for Complex Adaptive Systems

Developmental Evaluation is especially appropriate for situations with a high degree of uncertainty and unpredictability where the purpose of the evaluation is to support development, adaptation, and innovation in a complex adaptive environment characterized by rapid change and dynamic system relationships among important influences and actors. (See Chapter 8, pages 277–290, for an in-depth discussion of *Developmental Evaluation.*)

Developmental Evaluation

Developmental evaluation is especially appropriate for situations with a high degree of uncertainty and unpredictability where an innovator or funder wants to "put things in motion and see what happens." Using the example of the leadership development

program, infusing a hundred highly trained and supported national leaders into the nonprofit sector fits these parameters. All kinds of things can happen; many unexpected results may emerge (and did in the actual initiative on which this example is based). Some form of real-time, open-ended tracking is needed to capture what emerges. This is especially appropriate for venture capital and highly entrepreneurial, seed money efforts where the strategy is to infuse people and resources to shake up a system, increase the rate and intensity of interactions among system elements and actors, and see what happens. As things start to happen, additional resources can be infused to support further development. The very act of capturing what emerges and feeding information back into the evolving system makes this form of developmental evaluation part of the intervention (a form of process use). In such developmental evaluation, there is not and cannot be a classic separation between the measurement process and what is being observed (Gamble 2007). The observations affect the emergent self-organizing by injecting new information into the dynamic and emergent system. For example, following up with the trained leaders to find out who they are networking with and what is emerging can stimulate them to network. Even the unit of analysis can change in emergent, developmental evaluation as the evaluator tracks whatever emerges; that is, the very definition of units of analysis is emergent as the evaluator follows the complex nonlinear dynamics of the connections and linkages among those exchanging information and engaged in some forms of self-organized collective action.

Attention to complex adaptive systems provides a framework for understanding such common evaluation issues as unintended consequences, irreproducible effects, lack of program fidelity in implementation, multiple paths to the same outcomes, "slippery" program requirements, and difficulty in specifying treatments (Morell 2005:72). Glenda Eoyang (2006a) has described how a complexity-based approach to evaluation was used in a large social services department with 3,000 employees at the county level to help integrate services and functions. The evaluation elucidated emergent "networks of meaning," supported new approaches to alignment in looking at how different units connected to each other, and tracked the evolution of language about what was happening, which shaped emergent meanings. Eoyang and Berkas (1998) have examined and generalized the particular contributions that the lens of complex adaptive systems can bring to evaluation. They concluded,

> The system-wide behaviors of a Complex Adaptive System (CAS) emerge over time. For this reason, the evaluation system should focus on developmental and emergent patterns of behavior that:
>
> Match the developmental stage of the system. . . . Consider the dynamical pattern that the system exhibits over time to design an evaluation program to capture the "differences that make a difference."
>
> Track patterns and pattern changes over time, rather than focusing exclusively on behaviors of specific individuals or groups. While it may be unreasonable to expect a particular path of development or a predetermined outcome from a CAS, emergent patterns of behavior can be expected outcomes. An effective evaluation system must be able to capture and report on these evolving patterns. (Eoyang and Berkas 1998: www.chaos-limited.com/CAS.htm; see also Eoyang 2007)

The Challenges of Establishing Causality

Theories of change are important in evaluation because causality is a central issue in making judgments about a program's merit, worth, or significance. The classic causal question, in all its simple brilliance, is: Did the program produce the desired and intended results? Or, to what extent can the observed outcomes be attributed to the program's intervention? Establishing causality becomes more complicated over longer periods of time, when there are multiple interventions occurring at the same time, influences flow in both directions because of feedback loops, and interdependent relationships in a dynamic system means that all kinds of interconnected elements can change at the same time. The notion of cause and effect can lose all meaning in highly dynamic and complex adaptive systems characterized by nonlinear patterns of interaction, where small actions can connect with other smaller and larger forces and factors that wander hither and yon, but then suddenly emerge, like a hurricane, as major change at some unpredicted and unpredictable tipping point. Evaluation methods, designs, and tools for measuring predetermined outcomes and attributing causes for simple interventions that can be depicted as linear logic models are of little use in tracking the effects of interventions in complex adaptive systems. Indeed, the imposition of such linear designs and simple measures can do harm by trying to control and thereby stifling the very innovative process being tracked. Different kinds of theories of change, then demand different evaluation approaches to conceptualizing and assessing causality (Rogers forthcoming), including the importance of developmental evaluation for use in complex adaptive systems. The challenge is to match the evaluation approach to the nature of the situation, the theory of change at work, and the information needs of primary intended users.

In all this, it is important to interpret results about causal linkages with prudence and care. In that regard, consider the wisdom of this Buddhist story.

> *One day an old man approached Zen Master Hyakujo. The old man said, "I am not a human being. In ancient times I lived on this mountain. A student of the Way asked me if the enlightened were still affected by causality. I replied saying that they were not affected. Because of that, I was degraded to lead the life of a wild fox for five hundred years. I now request you to answer one thing for me. Are the enlightened still affected by causality?*
>
> *Master Hyakujo replied, "They are not deluded by causality." At that the old man was enlightened.*
>
> —Adapted from Hoffman 1975:138

Causal evaluation questions can be enlightening; they can also lead to delusions. Unfortunately, there is no clear way of telling the difference. So among the many perils evaluators face, we can add that of being turned into a wild fox for 500 years!

Follow-Up Exercises

1a. Identify a program that has identifiable and measurable outcomes for individual participants in the program and where the length of the program is less than one year. Examples would be programs that treat alcohol and drug abuse, parent education classes, an agricultural extension program for farmers, a job

training program, and so on. Based on the written materials available on that program, construct a linear logic model and explain the steps in the model.

1b. Develop a logic model with the staff in an actual program. Describe how staff reacted to the process of developing the logic model.

2. Take a logic model and convert it to a theory of change by explicitly identifying the causal mechanisms and/or assumptions that explain why each step in the model should occur. For example, a simple logic model would specify that increased knowledge about the health risks of smoking would lead to reduced smoking (Exhibit 10.3). What predictive theory undergirds and explains the causal connection between knowledge and behavior change in this example? Pick your own logic model example and add the causal linkage explanations and/or validity assumptions that convert it to a theory of change.

3. Pick a highly visible political issue and depict it as a logic model. For example, what is the "logic model" that explains why torture is used to extract information from presumed terrorists (e.g., the Abu Ghraib torture controversy in the Iraq War)? Construct the logic model for some matter of visible public policy and then analyze and comment on the assumptions of advocates. Examples of topics: solutions to global warming; relief of Third World debt; more gender equity by educating girls in developing countries; banning gay marriage; stopping family violence; ending child prostitution; building a high wall between the United States and Mexico to stop illegal immigration; teaching young people to abstain from sex until marriage; cutting down old growth forests for timber; drilling for oil in the Alaska wilderness; or any current public policy controversy. Portray the proposed intervention and alleged solution or desired result as a logic model. Discuss and explain the "logic" of the model you construct.

4. Exhibits 10.8, 10.9, 10.10, and 10.11 portray different ways of looking at a prenatal program for pregnant teenagers. Reconstruct those exhibits as a postnatal program. The intended and desired outcome of a prenatal program is a healthy baby and healthy mother when the baby is born. The intended outcome of a postnatal program is a healthy mother and child 2 years after the baby is born. Reconfigure those four models for a postnatal program. Discuss and explain the changes.

5. Pick an intervention of some kind, for example, a program aimed at stopping people from talking on cell phones (mobile phones) while driving. Pick your own example. Using the distinctions between simple, complicated, and complex (Exhibits 10.13, 10.14, and 10.15), present and analyze that intervention from the perspectives of (a) a simple problem that can be solved through a linear model; (b) a complicated problem that requires understanding and changing system relationships; and (c) a complex problem that is best portrayed as emergent and uncertain and requires developmental evaluation. In essence, look at the sample issue or intervention through these three different lenses. Compare and contrast what each illuminates and makes possible—and the evaluation implications of each perspective.

Note

1. Reprinted from *Angels in America, Part Two: Perestroika* by Tony Kushner. Copyright 1992 and 1994 by the author. Published by Theatre Communications Group. Used by permission.