Some of the essential points of applying emergent realist theory to practice are sketched in this chapter. First a running conversation is offered to explain emergent realist evaluation; then an example pre-proposal highlights important aspects of practice.

A Realist Theory of Evaluation Practice

Melvin M. Mark, Gary T. Henry, George Julnes

> It is by imagination that we can form any conception of what are his sensations.
>
> Adam Smith, *Theory of Moral Sentiments* (2nd ed.), 1762

> It was just my imagination, once again, running away with me.
>
> The Temptations (Norman Whitfield and Barrett Strong, songwriters, 1971)

In this chapter and in those that follow, we present a realist theory of evaluation. Realism is a tradition of increasing importance in evaluation (House, 1991; Pawson and Tilley, 1997), and we refer to our approach as emergent realism. Shadish, Cook, and Leviton (1991) list practice as the last of the five components of evaluation theory; the other four components provide the *terra firma* on which practice is constructed. For example, the knowledge construction component (see Julnes and Mark, Chapter Two) serves as the foundation for choices about research methods. But we have chosen in this sourcebook to bring the practice component to the forefront. This provides the best distillation of emergent realist theory in that it both introduces and summarizes the other components. By beginning this sourcebook with the practice component, we hope to introduce many of the most salient characteristics of emergent realist evaluation (ERE) theory and to show how ERE differs from other theories of evaluation. This discussion of practice, however, is based on emergent realist principles elaborated more fully later in this volume.

To better form a concept of what ERE is, we ask you, the reader, to exercise your imagination by placing yourself in the role of an emergent realist evaluator and considering how an emergent realist views a variety of questions that arise in the practice of evaluation.

As you prepare to leave the house, you, the emergent realist evaluator, think about the meeting you have scheduled today with a fellow evaluator to talk about collaborating on a proposal for an evaluation of a state program for preschoolers. You think the two of you can use emergent realist ideas to frame the proposal, but you need to introduce your colleague to emergent realism. You think you might say something like. . . .

Emergent realist evaluation, or ERE, is an approach to evaluation based both on a philosophy called neo-realism and on an integration of lessons drawn by many evaluators from practice, theory, and research. Like other forms of realism, emergent realism assumes that reality exists apart from our understanding of it. Emergent realists believe, to take a timeworn example, that if a tree fell in a forest, its crash would create the sound waves that could be registered as sound, even if no one were there to hear it. While this might seem like an obvious assumption, it actually differs from the idealism held by some constructivists who deny that there is any reality apart from our interpretations. Emergent realism also assumes, as do most forms of realism, that there are regular patterns in the world that we can detect, including *underlying generative mechanisms.* When we observe an event—say, a tree falling in the forest—the event is caused by unobserved processes, such as gravity, that occur at other levels. Although gravity, as an underlying mechanism, is not itself directly observable, it has detectable consequences, such as the falling of a tree. Similarly, the crashing of the tree makes detectable sounds because of underlying regularities involving sound waves and their impact on the human ear. Unobservable underlying mechanisms give rise to observable events.

Reality, then, is hierarchical or stratified, in the sense that there are molecular, reductionistic levels (such as gravity and airwaves) and molar, holistic levels (such as trees falling and producing detectable sounds). In general, entities at the more molecular levels, such as gravity, are referred to as underlying generative mechanisms or structures, while those at the more molar levels, such as falling trees, are referred to as events and, when they are actually observed, as experiences.

There are other aspects of the emergent realist approach that it would be nice to convey, you think, such as the concept of "emergent properties" and "open systems," but this is already too deep for a morning meeting about writing a proposal. Maybe later. But you'd better mention, early on, a bit on the ERE perspective on sensemaking and valuing.

Emergent realists also assume that, as complex organisms in the world, humans have evolved sensemaking capabilities for understanding the world. For example, humans have evolved the ability to hear sounds that correspond to certain wavelengths. Unlike so-called naive realists, emergent realists assume that these evolved sensemaking capabilities, although adaptive, are imperfect—

they are indirect, constrained, and influenced by past experience. For example, one might "hear" a tree falling when no tree actually fell because of expectations, visual cues, and background noise. Emergent realists also recognize that, in addition to biologically evolved sensemaking processes, there are socially constructed sensemaking technologies designed to supplement the biological processes. For instance, one might use a hearing aid to better hear falling trees, conduct an experiment to better estimate the effects of some intervention, or carry out a survey to gather information from a larger and more representative set of individuals than one person could readily observe or interview. With critical realists, emergent realists bring a critical perspective to the assessment of the quality of these constructed technologies as an aid to evolved sensemaking. While constructed technologies *may* improve understanding, each one is fallible, and its limitations should be analyzed in the specific context in which it is used. The ERE perspective on evaluation methods, then, is that they are constructed technologies, to be used—with critical analysis of their value *in situ*—to help in making sense about social programs.

But ERE sees sensemaking as just one of the two prongs of evaluation. *(In fact, one key difference you saw between emergent realism and previous theories of evaluation—a difference that drew you to emergent realism—is its dual focus on sensemaking and valuing.)* Prior to ERE, most realist efforts emphasized only sensemaking. For example, Campbell, a sophisticated critical realist, emphasized the use of techniques designed to estimate the impact of programs. This is a form of sensemaking whereby the evaluator intervenes in the world, through the use of research design and measuring instruments, with the goal of better understanding aspects of the world (specifically, the program's effects). In contrast, some theorists, such as Guba and Lincoln (1989), focus more on addressing the value issues surrounding programs. ERE in a real sense represents a third alternative that focuses on both sensemaking and value-probing.

The emphasis on values in ERE serves partly as a counterweight to the inadequate attention historically given to values in the experimentalist tradition in evaluation, and partly as an explicit recognition of the value-laden context within which evaluations are conducted. For example, an evaluation of a social program or policy can rarely measure all of the possible consequences, intended and unintended, that each stakeholder group and the public values. But by measuring those outcomes that are related to some groups' value positions, an evaluation will legitimate those values, while giving short shrift to others. Emerging realist (ER) evaluators explicitly recognize that different individuals and groups assign varying levels of importance to different values, and that choices made in the evaluation process can serve some value perspectives and the parties that hold them over others. In addition, ER evaluators believe that the value positions surrounding social programs can and should be directly studied, and it is this belief that differentiates the ERE position on values from many other approaches to evaluation.

The issues concerning values are tricky, but enough for now. The general point you'll need to make up front is that ERE consciously addresses both sensemaking

activities, such as estimating the effects of programs, and valuing activities, such as assessing whose values are served by particular outcome patterns. And you know that your colleague is an experienced and savvy evaluator, who is well aware of the value conflicts that can arise in the design of evaluations and in the utilization of evaluation findings, so you expect her to be receptive to your points about values. But you'll have a limited amount of time to meet and discuss your prospective collaboration. And you expect that one of the first things she's going to ask, if you push your ERE perspective, is what difference ERE makes in terms of how you will frame the evaluation. So, as you drive to the meeting, you think about. . . .

How does an adherent of ERE specify priorities among the various purposes that evaluation might serve? Of the many purposes that evaluations might have, what is generally most important? And how might the particular circumstances of a given evaluation lead to different priorities?

Although a more nuanced and contingent realist view of evaluation is forthcoming (Mark, Henry, and Julnes, in preparation), realist evaluation to date (House, 1991; Pawson and Tilley, 1997) has given general priority to explanation, in the sense of identifying (1) the manipulable generative mechanisms that underlie program effects, and (2) the conditions under which these mechanisms operate, including (3) the types of individuals for whom they operate. This approach, if it works properly, serves three functions of evaluation: formative, summative, and knowledge construction (Scriven, 1990; Patton, 1997).

As you think about this point, you try to recall one of your favorite quotes from Lee Cronbach, who was one of the first and foremost advocates of explanation as a focus of evaluation. In the quote, Cronbach indicates the desired overlap of formative and summative evaluation and knowledge construction. If you could recall the quote verbatim, you would remember it as

> Evaluation that focuses on outcomes can and should be used formatively. When a trial fails, the social planner wants to know why it failed and how to do better next time. When the trial succeeds and the proposal is considered for use under changed conditions, the intelligent planner does not conclude that its effectiveness has been proved. She now asks about the reasons and essential conditions for the success. Even when the trial of a plan has satisfactory outcomes, the policy maker should be prepared to consider any alternative that has a chance of working appreciably better (Cronbach, 1982, p. 12).

So a focus on underlying generative mechanisms aids formative purposes by providing a basis for strengthening an intervention, making it more efficient, better targeting it to those who will be most helped, and designing future interventions—all of which are easier to do if you understand *why* a program works. Learning about underlying mechanisms also serves traditional objectives of summative evaluation. In particular, knowledge of generative mechanisms is a valuable basis for generalizations about the likely effectiveness of the program for other persons, settings, or times (Cronbach, 1982; Mark, 1990; Cook, 1993). In addition, studying underlying mechanisms requires assessing the effects of a program, usually on multiple outcomes. Thus, the emergent

realist's focus on underlying mechanisms will also typically provide information about the merit or worth of the program, in the sense advocated by Scriven (1990).

Another car cuts into your lane, causing you to turn your full attention to driving. As you get back to thinking about your meeting, you realize that your colleague will probably push you on ERE's focus on explanation. You can hear her ask: Should every evaluation focus on underlying mechanisms? Aren't there areas where we have social programs but lack the substantive knowledge needed to develop good explanations? And sometimes aren't there important evaluation tasks that should be done that don't involve explanation? You plot your reply. . . .

Although emergent realist evaluation places a priority on research that probes underlying mechanisms, it does not hold that *all* evaluation should be of this form. When interventions are new and rapidly evolving, especially when the intervention and/or evaluation resources are small, it may be preferable to do more formative evaluation work, such as evaluability assessment (Wholey, 1987). Moreover, some interventions are "puny," as illustrated by the one-hour speaker approach to diversity training or sexual assault awareness prevalent on some college campuses. Mechanism-probing evaluation research is typically labor and resource intensive. While it should be helpful to look at puny interventions through an ERE lens, it is probably not cost-effective to evaluate them with a full scale mechanism-probing study.

Moreover, explanation can sometimes take place at a relatively molar level. (See an extended discussion of this issue in Chapter Two.) While emergent realists emphasize the value of explanation, they also insist, first, that explanation occurs at different levels and, second, that a highly molar explanation, containing simply the description of the causal relationship, can be important and useful. Human history is filled with cases in which people recognized and used regularities without a satisfactory molecular explanation. Pre-industrial humans recognized many toxic substances and developed rules of diet and, in some cases, practices such as using poison darts for hunting, without an adequate explanation of the underlying physiological processes. People have long used electricity and aspirin without good explanations for their effects. Similarly, the harmful average effects of smoking have been recognized without a satisfactory account of the mechanism by which it operates. Indeed, individuals and social groups can be right about useful causal regularities while holding an incorrect explanatory account. For example, manipulation of the spine by a chiropractor is effective for lower back pain; fortunately, we do not have to believe chiropractic theory, which holds that virtually all medical problems arise from spinal misalignment, to take advantage of this effect (Carey and others, 1995). In short, although a good account of underlying mechanisms is valuable and often leads to more efficient or more effective action, it is also possible to have useful knowledge about causal regularities without having a molecular explanation. Sometimes knowing an average effect size, for the specific contexts in which a program operates, is good enough for decision-making for the time being, as manipulability theorists (Shadish, Cook, and Leviton, 1991)

such as Campbell have argued. Thus, though emergent realist evaluators aspire to good molecular explanations as a goal of evaluation, they also believe the position held by manipulability theorists can be a useful fallback, at least if one can be confident about the applicability of the findings in the particular context(s) of interest. In short, when the level of knowledge is not adequate for testing explanatory accounts of underlying mechanisms, ERE endorses research that (1) allows more molar causal assertions about the effects of a program, with emphasis on identifying the conditions under which particular effects occur, and (2) helps develop knowledge about underlying mechanisms.

Viewed from a developmental perspective on knowledge construction, an emergent realist approach to evaluation involves seeking a relatively complete account of underlying mechanisms, but with the realization that such accounts may *emerge* over time and research efforts. ER evaluators should not, however, require or expect a detailed, empirically supported explanatory account for all actions or all recommendations (see, for example, House, 1991, on the Mackie's analysis of causal relations). Good evidence about program effects may suffice for judgments of merit and worth, even if explanation is molar or lacking. But less confidence will be warranted in generalizations to other sites or in formative suggestions about program improvements, relative to the degree of confidence one could have if better knowledge about mechanisms were available.

You think about making one other point about those evaluation activities that do not involve explanation: Contemporary realists are concerned not only with underlying generative mechanisms, but also with underlying structures or categories. Many evaluation tasks can be analyzed in terms of this focus, but the development of ERE to date has emphasized sensemaking with respect to underlying mechanisms (see Chapter Two; for a more comprehensive account of realist evaluation, including a description of classification and other evaluation tasks to carry out when explanation is of less interest, see Mark, Henry, and Julnes, in preparation). "So enough about explanation and sensemaking already," you imagine your colleague saying, "What about this values piece of ERE you mentioned?"

In addition to the more familiar knowledge construction purposes, a major purpose of ER evaluation is to unpack the value issues surrounding a program. ERE sees values-probing as a core focus of evaluation activity, in conjunction with the sensemaking activities that center on estimating program effects, assessing contextual determinants of success, and probing underlying mechanisms. Values are important throughout the life cycle of programs, from the definition of a social condition as a social problem, to the selection of program alternatives, to the decision about whether to initiate a particular program, to decisions about whether to continue, expand, or modify a program. Values influence evaluation use. Entrenched—but typically implicit—value positions in the "policy setting community" (Cronbach, 1982; Kingdon, 1995) can prevent evaluation from sparking program improvements. In part this occurs because stakeholders may draw conclusions about programs based on

value positions that are relatively resistant to information about effects (Ross and Nisbett, 1991). For example, in the treatment of homeless individuals in the United States, some people are committed to the building of dwellings in residential neighborhoods not because of any evidence that these result in better outcomes, but because they see it as a matter of human dignity and decency. Others respond, "Not in my backyard" to the construction of shelters, regardless of any evidence about the impact of such shelters on crime, property values, or any other outcomes (Wright, 1991). In short, evidence about program effects, even evidence about underlying mechanisms, may fail to influence the conclusions stakeholders will draw because values can overcome the effect of evidence. In addition, those involved in debates and decisions about programs (elected officials, for example) are often influenced by assumptions about the values held by others (such as constituents). Thus, evaluation can inform policy processes by addressing such questions as "What sort of outcome pattern satisfies whose values?"

Unlike Guba and Lincoln (1989), ERE does not view the evaluator as the necessary vehicle for negotiating value discrepancies. But ERE does acknowledge that policy-making and social programming are not based solely on information about underlying mechanisms. Evaluations may be more potent—at least in the sense of stimulating discourse—when they also provide information about which outcomes, generated by which mechanisms, are valued by whom. Of course, as with other social phenomena, the emergent realist evaluator recognizes that values may be affected by context and are susceptible to change over time—which may have implications for the assessment of values and the reporting of findings about them.

As you make the last turn approaching the meeting site, you imagine your colleague's next comment. You know she's read Shadish, Cook, and Leviton (1991) and will point out that, even when the general purpose of an evaluation is determined, choices are likely to remain about which specific questions to address. If, for example, one's purpose is to estimate the effects of a program, the issue of question choice arises in terms of how to decide which of the many possible effects one should measure. You think about this as you approach the building.

Question choice for the mechanism-probing brand of ER evaluation is, in one sense, quite simple. Evaluators should address those research questions that help identify which mechanisms are operating and which are not. For example, if the addition of a given dependent variable will help differentiate between two possible mechanisms, it is important to add measures of this variable. For instance, Entwisle (1995) indicates that, in research on the long-term social effects of preschool programs, one possible mechanism is that preschool participation allows some of the participants to avoid negative tracking, such as placement in lower-level reading groups or retention in the early grades. In evaluating a preschool program, then, this potential underlying mechanism could be probed by assessing the impact of preschool participation on tracking assignments. Question choice is based on program theory, then, and the

evaluator will typically focus on estimating effects relevant to program theory, as well as on testing potential moderators and mediators of these effects. For example, the effect of preschool could be moderated by the quality of the schools the preschool participants enter (Lee and Loeb, 1995).

However, other criteria may also often be needed to guide question choice. In some cases, for example, only a few outcome measures will be needed to differentiate among alternative mechanisms, and another criterion will be needed to select among the other possible measures. In other cases, the knowledge base in a particular area may be inadequate for a strong, molecular search for underlying mechanisms, and some other criterion may be needed to select among possible measures. On what basis, then, other than importance for testing mechanisms, should the evaluator select among the many possible effects? The ERE answer, put simply, is that the evaluator should select, first, those effects that reflect the public's interest in the program and, second, those that are important to other important stakeholders.

For ER evaluators in particular, another issue arises with respect to question choice: At what level of molecularity should we seek explanatory accounts? Explanations for a social program's effects can be given at different levels of analysis, from the neurological to the psychological to the social-structural. From this perspective, ERE replaces the metaphor of the program as a black box to be explored with the more apt metaphor of a set of Russian stacking dolls, where another level of explanation always underlies whatever level we are examining. ERE does not call for evaluators to move inexorably on to more molecular levels of analysis. While such an approach might be appropriate for basic research, it is impractical in the applied world of evaluation. Instead, utilization serves as a practical determinant of how far to go within the set of nested dolls. For example, there would be no reason to move from a psychological to a neurological level of analysis, unless doing so could reasonably lead to better agenda setting, selection of alternatives, choice of an alternative, or implementation activities (see Henry and Rog, Chapter Five).

A similar concern arises in terms of how far evaluators should go in looking for additional contextual determinants of program success. Testing alternative mechanisms often involves looking for moderators, that is, for contextual and client factors that change program effectiveness. (In statistical terms, this involves testing for significant interactions between the program and client or contextual factors). For example, a treatment program for homeless substance abusers might be found to be ineffective for homeless who have a psychological disorder, but effective for others. But a focus on client and contextual boundaries can lead researchers into Cronbach's (1975) "infinite hall of mirrors," whereby we encounter ever higher-order interactions that limit the initial lower-level contextual interactions. For example, the interaction between the homelessness intervention and client psychopathology might take place in a higher-order interaction, such that the treatment is effective for homeless without a psychological disorder only at sites where other social services are

adequate. The molar-molecular dimension also applies to the study of interactions. The focus on increasingly higher-order interactions can be seen as molecular—at its extreme, focusing on the response of each individual in a single setting—relative to the molar main-effect approach involved in examining only the average effect of the treatment relative to the control group. Some scholars (such as Guba and Lincoln, 1989) have reacted to the possibility that a finding might be limited by higher-order interactions by insisting on highly molecular analyses and, further, by denying the utility and even the existence of social regularities. ERE is not sympathetic to this stand which, as its critics note, is embedded in a radical relativism which leaves us with no warrant for any action (for example, see Pawson and Tilley, 1997). The question remains, though, of how to choose the level of molecularity to be employed in our explanatory accounts. ERE suggests that the level of molecularity should be driven by (1) utility, so that explanations are at a level corresponding to the needs and conceptions of those who will use an evaluation, and (2) empirical evidence, regarding the large (and policy-relevant) contextual determinants that need to be accounted for so that the explanatory account is a useful guide to practice.

You park your car, grab your briefcase, and head inside. After greetings, the meeting begins. To your surprise, it begins pretty much as you expected. The one thing you hadn't had time to think about in advance, of course, turns out to be the one thing that your colleague is really interested in—the issue of "method choice." That is, what guidance does ERE provide on how to select the research methods for a particular evaluation? In trying to be concise, you discuss this issue a bit mechanically.

The primary criterion for method choice for sensemaking, according to ERE, is the method's suitability for probing underlying mechanisms. ER distinguishes between two general approaches to the study of underlying mechanisms: (1) *competitive elaboration,* which can be applied when alternative generative mechanisms are articulated in advance, and (2) *principled discovery,* which can be applied when generative mechanisms are not adequately specified in advance. In fact, competitive elaboration and principled discovery share the same underlying logic, but the distinction is useful in practice.

You decide to explain competitive elaboration first, both conceptually and in terms of methods that can be used to carry it out, even though principled discovery might be carried out first. Why hadn't you thought about method choice on the way over?

Competitive elaboration refers to the process by which alternative explanations—whether alternative program theories or validity threats—are ruled out (Reichardt and Mark, 1998). Competitive elaboration involves (a) specifying the implications of a possible mechanism to discover those that conflict with the implications of alternative mechanisms (including validity threats), and (b) obtaining data to see whether the implications of the mechanism or of its competitor hold true. In other words, when an explanatory account is susceptible to alternative explanation, the plausibility of the alternative—and of

the original—can be put to test by adding a comparison that creates competition between the explanatory account and the alternative explanation. The comparison can be made with respect to any of the elements of a causal relationship: cause, recipients, setting, time, and outcome variable (Reichardt and Mark, 1998). For preschool programs, Entwisle has articulated three potential generative mechanisms for the long-term effects that have been observed (1995; for descriptions of the effects, see Barnett, 1995; Consortium for Longitudinal Studies, 1992; McKey, 1985). The first possible mechanism is that increases in IQ score allow preschoolers to avoid negative tracking assignments, and the better tracking assignments in turn stimulate better outcomes. A second potential mechanism is that preschool participation causes parents, teachers, and others in the child's social world to hold higher expectations for the children, and the children respond to the enhanced socialization that results. According to the third of Entwisle's possible generative mechanisms, the preschool serves directly as a trigger to make the children ready for the rigors of kindergarten and first grade. Each mechanism specifies some short-term changes that can be measured and tested. In addition, an ER evaluator would be aware that each mechanism might work differently with children from different family and social environments. (For other examples, see Mark, 1990; Mark, Hofmann, and Reichardt, 1992; and Reichardt and Mark, 1998).

Competitive elaboration can be accomplished through a variety of research designs: (1) In the quantitative domain, moderated multiple regression, analysis of variance, or other techniques can be used in planned analyses to assess causal mechanisms by determining whether observed program effects are robust across different client subgroups, settings, outcome variables, treatment variations, and time lags. For example, in a well-known time series quasi-experiment, Ross, Campbell, and Glass (1970) found that a crackdown on drunken driving in Britain was followed by reduced traffic fatalities during the hours pubs were open, and that no such decline occurred in hours when pubs were closed. This pattern helped demonstrate that the crackdown reduced drunken driving, and it helped rule out a number of alternative explanations. Analyses that assess the robustness of program effects across subgroups, settings, outcomes, or time, even if not definitive in pointing to one mechanism over others, can be useful to policy-making in terms of showing possible limits to generalizability. (See Julnes, 1995; Mark, 1990; Mark, Hofmann, and Reichardt, 1992; and Reichardt and Mark, 1998, for examples and discussion of these and other benefits of this approach.) (2) Hierarchical linear modeling (HLM) and related techniques can be employed to study the effects of variables at different levels of aggregation (Bryk and Raudenbush, 1992). These techniques have been commonly used in education, for example, to examine the interactions of school-, classroom-, and pupil-level variables. HLM can be used, for instance, to investigate community-level effects and how they interact with program effects. (3) Meta-analyses can also provide evidence about moderated effects when a number of evaluation studies have been conducted (Cook and others, 1992; also see House, 1991, who argues

for meta-analysis from a scientific realist perspective). (4) We can also learn about mechanisms through the study of mediation, which is another form of competitive elaboration whereby one's program theory predicts that the program's effect on an outcome will be mediated by some other variable, whereas alternative explanations do not make this prediction. For example, Seligman's revised theory of depression predicts that those in therapy will first show a change in the sort of causal attributions they give for negative events, and that the use of these attributions will statistically account for subsequent change in depression; alternative explanations do not make this prediction. (5) In some instances, as Pawson and Tilley (1997) point out, basic research will be informative about generative mechanisms in social programs. For example, Borkovec and Miranda (1996) use the results of laboratory experiments from cognitive psychology and other areas as part of the web of evidence to argue for their model of treatment for generalized anxiety disorder. (6) Qualitative methods can also be employed for competitive elaboration (Yin, 1994; Campbell, 1974; Maxwell, 1996; Mohr, 1995). For example, in a recent evaluation of the impact of a technology funding program on instructional processes, through case study methods a few specific mechanisms were identified as necessary but, by themselves, insufficient to stimulate an impact (Jones, Dolan, and Henry, 1996). On-site teacher training and availability of technical support for labs and classroom equipment, for instance, were found to be present in many situations where technology was most highly used. However, each component was also present in cases where the technology was not heavily used. The generative mechanisms underlying high use of technology were observed to take a varied, but not entirely unsystematic, route. In a follow-up study, ten sites are being used as case studies to further elaborate these mechanisms and to see if there are discernible packages or configurations that seem to lead to higher levels of instructional use and under what circumstances these configurations work. Smith (1997) presents a similar example that relies on the integration of qualitative and quantitative research. (7) More generally, competitive elaboration involves the use of pattern matching (Campbell, 1966; Trochim, 1985)—which can in fact encompass all the preceding approaches. For example, in an evaluation of Georgia's HOPE (college) scholarship program, Henry and Bugler (1997) hypothesized that the program's requirement of a B or better average in college and its limit of four years of support would serve as potential triggers for several outcomes: higher GPA, more credit hours earned, and greater likelihood of remaining in college after two years of study. Program impact was tested on a sample of marginally qualified students and on subgroups of traditionally underserved students, specifically African Americans and women. The HOPE scholars earned more credit hours, had higher GPAs, and were more likely to remain in college. However, this was true whether or not the HOPE scholars had retained their scholarships. This pattern of results is consistent with the interpretation that, underlying the HOPE program's effects, is a mechanism whereby it legitimizes scholarship recipients as college students (Gamoran, 1996).

"So," your colleague asks, "for competitive elaboration, you start with a program theory—or at least a hypothesis about an underlying mechanism, right? But what does ERE say you should do when you don't start with a good hypothesis about the underlying mechanism?" Ah, you reply, that's precisely what "principled discovery" is about—how to do elaboration in the absence of strong theory.

In many cases in which evaluations are done, theories may be inadequate. And having 100 equally speculative theories is probably no better as a guide to research than having none. Moreover, in some cases, programs are evaluated before practitioners are able to develop the experientially based theories that can guide realist evaluations. What do we do in these cases? How do we ask the data, rather than practitioners or social science theory, to provide the program theory to further guide us? There are several possible approaches. (1) The exploratory data analyses of Tukey (1977) and other graphical methods (Henry, 1995) could be used more widely to guide informed speculation about underlying mechanisms. Even if not predicted in advance, the observation that larger effects cluster in one subgroup, or in one setting, should set off additional investigation and the search for the underlying mechanism. (2) Techniques such as regression and analysis of covariance can be used in an exploratory fashion—and indeed may be most important as exploratory tools (Tukey, 1977)—to search for variations in treatment effectiveness across subgroups and settings. One approach is to use residuals to identify sites or cases that have larger or smaller outcomes than would be expected from standard predictors. These extreme cases can be contrasted to see if the variation in outcome appears to be associated with differences in types of participants or in treatment implementation, relying perhaps on qualitative data from interviews or observations. However, these and other efforts at discovery should be principled—-for example, evaluators should not present an interaction they stumbled upon as though it were the confirmation of a hypothesized mechanism. Rather, such discoveries should lead to theory building and be subjected to replication, other tests, or both.

(3) An important form of principled discovery is what Julnes (1995) calls the "context-confirmatory approach." Under this approach, an empirical discovery that suggests a mechanism (such as the discovery of differential effects across subgroups on a key outcome) is used to generate a distinct prediction that should be true if the newly induced mechanism is operating (such as the pattern of differences in some other outcomes). (4) It is even possible to obtain evidence of differential effects, without certainty about their origin, by testing for variance shifts that result from the program. Techniques such as those described by Bryk and Raudenbush (1988) can be applied to assess whether treatment groups differ in variability in such a way that would indicate that some unknown interaction exists. For example, if the variability among those assigned to workfare is greater than among those assigned to traditional welfare, this could indicate that workfare is (relatively) helpful to some and harmful to others. Such knowledge (a) could lead to cautions about proposed policy changes and (b) should initiate an exploration to identify what differentiates

those who are helped from those who are hurt. In short, emergent realist evaluators have a considerable array of methods available with which to interrogate the data, which is particularly important when pre-existing theories are not sufficient to guide the search for mechanisms and for the contextual determinants of program outcomes.

"Okay," says your colleague. "I like competitive elaboration and principled discovery as two general strategies. But you mentioned a lot of different ways you could do each of them. How do we decide which specific method to use?"

The concepts of competitive elaboration and principled discovery themselves do not lead magically to method choices. Different methods can be used to carry out competitive elaboration or principled discovery, and the evaluator needs to make an informed choice of methods for the particular evaluation context. With respect to the question of whether to adopt the methods of competitive elaboration or those of principled discovery, the state of existing knowledge will be the primary determinant. The less is known about possible mechanisms, the more likely the evaluator will need to employ the more exploratory techniques of principled discovery, rather than the more verification-oriented techniques of competitive elaboration. But, as suggested by the context-confirmatory approach (Julnes, 1995), the ERE evaluator will probably use both, sometimes in an overlapping fashion and sometimes iterating between the two.

"Let me try this out," interjects your colleague. "Let's say you're evaluating a boot camp program for criminal offenders. You start with a more exploratory analysis, like stuff Tukey wrote about, right? Let's say you find that, compared to the traditional criminal justice system, the boot camp program reduces recidivism for offenders with minor criminal records but not for offenders with more severe records. From that finding, you generate the hypothesis that the mechanism underlying the program is 'labeling': The boot camp keeps minor offenders from getting labeled, by themselves and by others, as criminals." You start to respond, but your colleague continues: "And so from this potential mechanism you develop another hypothesis, such as that, controlling for offense, the program will be more effective for younger than for older offenders because the younger ones will be less likely to have strong labels as criminals." Exactly, you reply, except that the "competitive" part of competitive elaboration kicks in. You also need to identify any alternative mechanisms—whether drawn from the literature, program staff, or clients—such as whether the effects you've mentioned might be caused by motivation or by increased work skills instead of by labeling. And then you try to find predictions that will tell these apart. For example, if it's labeling, the difference between younger and older offenders presumably should be the same at all sites. But if it's work skills, the age difference is likely to depend in predictable ways on the specific activities carried out at each site to teach work skills. After you and your colleague congratulate each other for being able to talk the same language, you decide that you had better clarify the typically iterative relationship between principled discovery and competitive elaboration.

If you start without a strong theory about possible mechanisms, you begin with principled discovery and move on to competitive elaboration as much as

possible. On the other hand, if you can identify the possible mechanisms in advance, you start with competitive elaboration and work to collect observations that will help indicate what mechanism is in operation. But you can still do the more exploratory sort of work labeled principled discovery. This may lead to a refined theory of the mechanism (for example, if you find that the relationships predicted by the mechanism hold in some conditions but not in others) or to exploring a possible mechanism you had not identified in advance.

Your colleague jumps in: "Don't I remember seeing something on EvalTalk, or maybe in a review somewhere, that realists don't believe in control or comparison groups? What's this about? Isn't an experimental design going to be helpful sometimes in sorting out competing mechanisms, as well as in just figuring out what the effects of a program are?" I know what you're referring to, you reply. From the ERE perspective, it's a serious flaw in a book that has many other nice attributes.

Pawson and Tilley (1997), working from a scientific realist perspective, have argued against the value of experimental and quasi-experimental control groups. For example, in reviewing one evaluation, they note the study was "a quasi-experimental one, comparing changes between the experimental and control [sites]. However, . . . this cannot add anything instructive, and indeed it does not do so" (Pawson and Tilley, 1997, p. 97). Pawson and Tilley suggest that the "realist experiment" involves creating the conditions that allow one to observe whether a particular mechanism is triggered, and that experimental and quasi-experimental designs are of little value in doing so. On the latter point, ER evaluators disagree. Bhaskar (1978, p. 53), a key figure in critical realism, claimed that there are two functions that the experimental scientist must perform: "First, [s]he must trigger the mechanism under study to ensure that it is active; and, secondly, [s]he must prevent any interference with the operation of the mechanism. These activities could be designated as 'experimental production' and 'experimental control.'" The value of the randomized experiment, as Campbell (1957) suggested, derives from the value of the randomized control group in preventing (or, more precisely, accounting for) the interference of plausible extraneous mechanisms, such as generative mechanisms sparked by other events (history), or emanating from naturally occurring processes within the research participants rather than from the program (maturation). So the ER evaluator sees value in traditional experimental designs, to the extent they aid in principled discovery or competitive elaboration in a given instance.

"Um, where exactly do values fit in here?" your colleague asks, with furrowed brow. Well, you answer, we've kind of ignored values momentarily, but. . .

The sensemaking and values-probing in ER should be linked in an iterative process—a sort of ongoing churning in evaluations and in the policy-making community (Majone, 1988) more broadly. For example, one relatively likely form of this iterative and overlapping focus on values and sensemaking begins with values probes that provide an underpinning for the collection of

evidence for sensemaking, which itself will likely entail reinterpretation of evidence and the collection of additional evidence, which is interpreted in light of critical assessment of the values implicit in the evidence, and so on. In addition to feeding into the design and interpretation of the evaluation, findings about values should be communicated to the multiple audiences of the evaluation. While clarity about values conflicts will not lead magically to consensus, it should generally improve the quality of debate and deliberation. To accomplish this interlinking of sensemaking and values-probing, specific approaches or methods for values-probing are required (in addition to the methods for principled discovery and competitive elaboration, which aid sensemaking about program effects and underlying mechanisms). ERE includes three general methodological approaches for values-probing, the purpose of which is to understand the consensus and conflict around values issues among the various groups that have a role in the pluralist, democratic policy processes (Henry, 1996).

The first method involves using sample surveys to better understand the magnitude of concerns about social problems, the perceived need for social or governmental intervention (or for changing an existing intervention), the acceptability of different types of interventions, and what respondents value among the different outcomes an intervention might achieve. For example, in an evaluation of a preschool program for four-year-olds, one might first identify key stakeholder groups, such as teachers, parents, and program administrative staff. In addition, ERE calls for including "the public," which can be represented by a sample of residents in the area served by the program. Each of these groups can be surveyed with two sorts of questions. The first would ask about the extent to which respondents would consider the program a success if it achieved each of a number of different outcomes. Responses to these questions create a *values map,* allowing us to specify whose values are served by particular outcome patterns. A second set of questions can locate the program itself, apart from its observed outcomes, in a values context. For the preschool program, values related to the nature of services and service delivery can be as important as the lists of desired outcomes. For example, one key issue is whether the target population should include only children from "disadvantaged" households (an equity concern) or, alternatively, whether universal access should be provided to this developmental program without targeting specific groups (an equality concern). Similar choices are at the heart of many values issues in American society (Yankelovich, 1994). Other value issues are also of interest, such as parental choice in selecting a preschool program for their child, and the decision about what type of organization should be allowed to provide the services (public, private, or not-for-profit). Survey data on public and stakeholder group views on such questions help in clarifying whose values the intervention represents.

A second approach to values-probing involves the use of any of a number of qualitative methods. Focus groups and intensive interviews can be used to

assess the value structure of various stakeholder groups, including the public. Ethnographic observation might also be used to gauge the value positions of program staff and clients. As an example of a qualitative approach to values-probing, Caracelli and Greene (1997) describe an evaluation by Greene and her colleagues of a program that included the elimination of tracking in a school district's science program. This intervention was viewed by minority parents as a way of furthering their children's education, and as the right thing to do. Many other parents saw it as a threat to the quality of their children's preparation for college. Qualitative and quantitative methods were used to "represent pluralistic interests, voices, and perspectives better and, through this representation, both to challenge and transform entrenched positions" (Caracelli and Greene, 1997, p. 29).

A third approach to values-probing falls under the category of *critical analysis*. ER evaluators think of this as a broad category, which can involve examining the value bases of a program (1) from a philosophical perspective on equity, equality, freedom, or other core values, as when House (1988, for example) critiqued programs from the perspective of Rawls' theory of justice, (2) from the perspective of a social philosophy, as illustrated by critiques of social programs from the perspective of feminist theory, or (3) based on a particular methodology that includes critical interpretation of values positions, such as critical discourse (White, 1994). The methodology of critical analysis is designed to expose the biases that lurk in the unthinking acceptance of social values: "[Critical analysis] is wary that any consensus will reflect the interests of those with institutional power in a society and probes for ways to challenge the dominant norms in a society" (White, 1994, p. 519). ERE does not require that critical analyses of these sorts be done, but neither does it discourage them. Such analyses can be revealing about the value issues that drive decisions about a program. However, critical analyses are not a substitute for direct investigation of the value positions of key program stakeholders and the public. Nor is it a substitute for sensemaking activities about the program's effects and underlying mechanisms, although it may lead to a critical examination of the information obtained through such activities.

As with sensemaking, the values-probing component of ERE will often be best served by the use of multiple methods. For example, a sample survey approach may give breadth, in the sense of the ability to generalize about the public's and other stakeholder groups' values (and to compare the values of different groups), whereas focus groups may give depth, in terms of being able to probe deeper into such issues as the reasoning behind particular value positions.

At this point, your colleague interjects, "Hey, does that mean that ERE has a stance on mixing qualitative and quantitative designs? You know how Greene and Caracelli (1997) wrote about different stances on mixed methods?" Well, you reply. . . .

ER evaluators take a positive stance toward mixed-method designs. As Greene and Caracelli (1997) point out, some researchers have claimed that dif-

ferent inquiry paradigms are incompatible and cannot be crossed (for example, see Guba and Lincoln, 1989). Others see the relationship between paradigm and method as weak, and argue for a pragmatic approach to mixing methods (for example, Patton, 1988). The ER position is that the historical paradigms associated with quantitative and qualitative inquiry are not inherently necessary, and that we should not be bound by them in making method choices. Emergent realism shares some similarities with the positivist tradition and some with constructivism, but it remains as an alternative to either. (See Chapter Two of this volume for more detail). According to ERE, the primary considerations that evaluators should use in choosing methods are the ability of the method, in the context of that particular evaluation and its constraints, to (1) further competitive elaboration or principled discovery, as appropriate, (2) clarify the values structure surrounding the program, and (3) address the desired levels of molecularity-molarity. ER evaluators recognize that all of our methods are constructed sensemaking techniques, and all are fallible. In this light, these objectives will often best be served by a combination of quantitative and qualitative methods.

Moreover, when quantitative methods are used, they need to remain flexible for changing questions and for discovery (Mark, Feller, and Button, 1997). Smith (1992; 1997) has similarly advocated flexibility in allowing questions to emerge, although the ERE focus is primarily on the emergence of understanding about mechanism. One strategy that maximizes flexibility is to plan evaluations as a set of smaller, interrelated and sequenced studies, rather than a single large study with a single design (Cronbach, 1982).

"Makes sense to me—although I may have to think a bit more about the paradigm issue. I take it that realism is being proposed as an alternative paradigm. But I don't want to talk about that—I get a headache if I think about 'paradigms' and philosophy of science for too long. How about something more down to earth, like utilization?" Good idea, you say, we need to talk about utilization at some point.

The dual focus on sensemaking and values-probing in ERE exists in light of the corresponding processes that occur in the policy-setting community. For example, in the area of homelessness, sensemaking is involved in such questions as "What is the magnitude of homelessness and who is affected by it," while values are involved in such questions as "What role does society have in preventing or ameliorating the homelessness that exists?" According to ERE, utilization occurs when the sensemaking and values-probing activities of evaluation influence the sensemaking and valuing of the policy-setting community. This can occur at any stage of the policy process (see Chapter Five of this volume).

As suggested in the earlier discussion of evaluation purpose, ERE views explanation—the identification of underlying mechanisms and the conditions under which they are triggered—as an important route to utilization. But it is not the only underlying mechanism for utilization. For example, an evaluation using a randomized experimental design might show that a program has no

net effect, but it might provide no information as to why the program did not work. Nevertheless, the results could be used to inform members of the policy community about the existing program and stimulate discourse about changing the program.

Whether or not it is based on explanatory mechanisms, utilization requires generalizing. Inferences are drawn from the concrete observations of an evaluation to other instances where the results will be applied: to the future, certainly, and typically to other settings, to different clients, to variants on the program. Mark (1986) suggested there were three principles that underlie such attempts to generalize (compare to Cook, 1993). The first is the "similarity principle," which involves generalizing to similar instances. This principle is illustrated by Cook and Campbell's (1979) "modal instance" approach to external validity and Cronbach's (1982) recommendation to analyze subsets of the data that correspond most closely to the cases to which one wishes to generalize. Such analyses, along with careful consideration of the similarity between the context of evaluation and that of the desired generalization, can suggest how reasonable it is to generalize in the absence of a sound explanation. A second approach involves the "robustness principle," which involves generalizing based on the finding that the program has had similar effects across diverse contexts (Mark, 1986). The robustness principle underlies Cook and Campbell's (1979) call for "deliberate sampling for heterogeneity" and Cronbach's (1982) similar call for an "extreme groups design" as a way to increase external validity. Emergent realist evaluations will typically assess how robust a program's effect is by testing for interactions with a range of contextual and client variables in the course of competitive elaboration and principled discovery. The third principle that underlies attempts to generalize, according to Mark (1986), is "explanation," which of course is the primary focus of most ERE sensemaking efforts.

You then mention that other specific strategies, such as linking an ERE evaluation with an ongoing monitoring system, such as an MIS, may be useful for making and testing generalizations (Mark, 1995), but that these ideas are tangential to your current focus. "When," your colleague grins, "have we ever let the fact that something is tangential stop us?" It might be a good idea to cut our tangents short, you suggest, at least until after we have done a preliminary project budget. Meanwhile, back to utilization.

One aspect that influences utilization, and is often inadequately discussed in the evaluation literature, involves how to communicate our evaluation results. While a variety of approaches can used in briefings and written reports (for example, see Hendricks, 1994; Patton, 1997), the ER evaluator is especially concerned about communicating the findings to various potential users, including the public. Informing the public necessarily requires concerns for obtaining the information through the existing media channels and working with journalists (Henry, 1996). Graphical displays may be particularly useful (Henry, 1997). For example, for a project evaluating school quality, Henry and others (1996) developed easily interpreted graphical displays, using one to five

stars to compare a school with similar schools on each of a set of performance dimensions, and up or down arrows to show the one-year trend in performance.

In addition, it may be useful to attempt to find ways to *socially signify* some results, that is, to describe the magnitude of the impacts in terms that have meaning for the audience. For measures such as recidivism or high school graduation rates, this may not be necessary, because the measure is already in a familiar and commonly understood metric. But the public and other stakeholder groups are often unfamiliar with the meaning of the measures used in evaluations. In such cases, research that translates the measure into more familiar terms may be helpful. For example, in an evaluation of an intervention designed to reduce the time required to hire federal employees, Mark, Feller, and Button (1997) surveyed managers about how much more quickly people would have to be hired to make a real difference in their work units. Sechrest and Yeaton (1981) have made interesting suggestions about how to socially signify results, but it remains an area ripe for future work.

"You know," your colleague says with a big smile on her face, "we could keep on talking about stuff all day, but I think I'm sold. From what I've heard, I like this emergent realist approach. It seems to capture in a coherent way a lot of best practice in evaluation and to add some important features. Let's do this proposal and try to do it from an ERE perspective." You gladly agree, and spend the next weeks stealing time to work on the pre-proposal (the granting agency has solicited short pre-proposals and will solicit complete proposals from a subset of those who submitted pre-proposals). After a couple of revisions, it looks something like this:

Evaluation Plan for the Georgia Pre-Kindergarten Program

Pre-proposal: draft #3

Georgia's Pre-Kindergarten (Pre-K) program is a full-day, developmental preschool program offered to all four-year-olds in the state. Funded by the Georgia Lottery, the program enrolled approximately 62,000 (about 60 percent) of the four-year-olds in the state in the 1996–1997 school year. Explicitly stated program objectives range from improving the educational and social outcomes of the children and their families to increasing human capital and lowering crime rates. The program also appears to have several more implicit objectives, such as reducing the gap between more and less advantaged children, fostering elementary school reform in the state, and engaging parents in actively making educational decisions for their children.

The purpose of this evaluation is five-fold: (1) to understand the extent to which these valued outcomes are obtained by the program participants, their families, and the state; (2) to distinguish which of the plausible program-related mechanisms is most likely to have triggered the attained out-

comes; (3) to determine the specific services that trigger the mechanisms for specific groups of children; (4) to examine the extent to which other influences in the children's social worlds, such as parental activities and educational experiences in the K–12 environment, moderate and mediate the attainment of these outcomes; and (5) to probe the consensus or conflict in the public and various stakeholder groups about the values related to the program.

One focus of the evaluation will be on the characteristics of pre-K services delivered, including the type and integrity of the curriculum; teachers' philosophies and practices related to child-centered instruction; quality of the facilities, resources, and instructional program; and organizational factors. These program characteristics are viewed as potential triggers that may initiate specific mechanisms which further a child's development. By relating actual services and specific attributes of services delivery to the attainment of socially desired outcomes by children from a variety of social backgrounds, the evaluation can provide information useful for many groups: for parents to use in choosing a pre-K site and in working with their children and pre-K staff; for program personnel in improving the delivery of services; for program administrators and policy-makers in making systemic changes in the program; and for the public as an informational base in the ongoing policy deliberations concerning the program.

The summary of the evaluation design presented in the next four sections should convey the scope of the proposed research; however, more detailed description of procedures and project management are left for the full proposal. We first describe the values-probing that will provide a mapping of the various outcomes that different groups, including the public, hold as most important for the program. Following that is a brief discussion of the mechanisms that may be stimulated by preschool programs and may underlie the valued outcomes attained by such programs. In the third section, we overview the proposed methods for assessing the program's effects and the causal packages that may lead to specific outcomes. Finally, plans are presented for disseminating the information gained from the evaluation to a variety of groups who may have use for the information.

Methods for Values-Probing. Much is expected of preschool programs in the United States and, more specifically, of Georgia's pre-K program. The importance of preschool programs has been highlighted by scholars of American political opinions and values, who have noted that programs such as Head Start and, by extension, Georgia's pre-K program, provide a mechanism for reconciling the core values of freedom and equality (Yankelovich, 1994). That is, the public holds in general high regard those programs that offer the prospect of creating a level playing field for all.

In the case of the Georgia pre-K program, legislative history and public debate, as well as prior research on related programs, make it possible to develop a reasonably comprehensive list of the possible outcomes of interest. Valued outcomes for the pre-K program range from those involving the par-

ticipating children, such as improving their cognitive ability and social outcomes, to those involving families, such as improving family circumstances and parental interactions with children, to broader societal benefits, such as lower crime rates and higher rates of educational attainment. In addition, while most potential outcomes derive from and are consistent with stated program objectives, some potential outcomes are not objectives of the program but nevertheless may be highly valued by some group. For example, by providing free day care, the program may allow families to reallocate some funds to other purposes. While not a goal of the program, this may be a highly valued outcome for some stakeholders.

There are several reasons to assess which values are held by which stakeholders. First, the study of values will inform the design of the evaluation, such as the selection of outcomes and mechanisms to be examined. An evaluation should address the question of the effectiveness of a program in terms that reflect the values of those concerned with the program, including the public, upon whom all governmentally provided programs depend for support (Kingdon, 1995). In this, as in most evaluations, resource constraints will presumably preclude evaluating the program's effects on *all* possible outcomes. Instead, because parents, families, program staff, program administrators and policy-makers, and the public all have a stake in the program, it is important to include their most highly valued outcomes. Second, findings about values will be useful in organizing reports about program effects. By knowing the value stances of each group, their perspectives can be better represented in interpretations of the findings. And third, results of the values-probing component of the evaluation will be communicated to the public and to various stakeholder groups, as described in the later section on reporting. Heightened awareness of the value stances of the various groups should enhance the quality of the discussion and debate about the use of other evaluation results.

Although in most past evaluations, any efforts to reflect stakeholders' and the public's values have been informal, two methods are proposed to systematically probe values in this evaluation. First, it is important to establish the outcome preferences of the public and various groups of stakeholders, that is, the pattern of program outcomes which they would consider a success. To obtain systematic information on the outcomes valued by parents, teachers, program administrators, and the public, we propose to conduct sample surveys with each of the four groups. Items will be developed to reflect various possible valued outcomes. These items will stem from a review of enabling legislation, prior public discussion of the program, the research literature, and from consultation with program experts and members of each group to be surveyed. The items will then be included in a self-administered questionnaire for the program administrators and in telephone interviews of the three other groups. For each group, the values will be ranked and analyzed for significant differences. In addition, the results will be compared across groups, and the underlying structure of the responses of the four groups will be analyzed using multiple group structural equation models.

The results of the surveys will subsequently be used to formulate scenarios that highlight trade-offs among highly valued outcomes. These will then be discussed in eight focus groups, with two groups drawn from members of each of the four stakeholder groups. The results will add depth to the survey results and provide information about trade-offs that might result from emphasizing one component of the program at the expense of others.

Potential Underlying Mechanisms. While some past studies show short-term effects of preschool programs on the cognitive ability of participating children, long-term social effects, such as decreased delinquency, higher high school graduation rates, and reduced assignments to special education and remedial education programs, appear to be more common than long-term cognitive gains (Barnett, 1995; Consortium for Longitudinal Studies, 1983; McKey, 1985). Research to date on preschool programs has led to the formulation of several general, alternative mechanisms that may lead to such positive social outcomes (Entwisle, 1995). (1) Preschool may cause short-term but significant cognitive gains, which in turn lead to better placements, such as fewer assignments to special education, remedial education classes, and programs for learning disabled students. By avoiding negative tracking that would otherwise occur (Alexander and Entwisle, 1996), the preschool children do better. (2) Preschool participation may increase the educational expectations of those in the child's social world (parents and teachers), thereby increasing the social resources available to the child and thus increasing the child's attainment. (3) Preschool may effectively promote the development of social skills and behaviors that influence the child's readiness for school as perceived by kindergarten and first-grade teachers. These social skills result in higher placements, higher expectations, and better social outcomes for the children. (4) In a slightly different vein, the preschool experience may be viewed as a part of a continuing educational process, where the gains from one point in the process are mediated by later processes. According to this mechanism, the gains from high-quality, developmentally appropriate instructional programs will be enhanced and maintained by complementary kindergarten programs, but dampened by kindergarten programs that do not have those attributes (Marcon, 1992, 1994; also see Lee and Loeb, 1995).

If one or more of these mechanisms is triggered by the pre-K program, the trigger is presumably some attribute(s) of the pre-K program as it is carried out. Consistent with this belief is evidence that the quality of preschool programs has implications for the cognitive as well as social development of the children, presumably by influencing the extent to which any of the long-term mechanisms are set into place (Marcon, 1994; Carnegie Task Force, 1994). Knowing which attributes of pre-K practice trigger which mechanisms could lead to the design of more effective pre-K programs and could inform parental choice about which pre-K site to attend. Therefore, we propose to examine the potential mechanisms for cognitive gains, enhanced social skills, and increased parental expectations and improved educational resources in the home. This

will involve (1) locating or developing, and then applying, a battery of measures of program quality at selected sites and (2) using the methods described in the next section to assess the extent to which any of the mechanisms (see 1–4 in the preceding discussion) appear to be operating. Note, however, that the enumeration of causal mechanisms and therefore the eventual measurement of the indicators of their operation may change, depending on what outcomes are found to be highly valued in the values-probing component of the evaluation.

Methods for Sensemaking. To better understand which of these mechanisms, if any, works for which groups of children and their families, we plan to use several methods for a strategy of competitive elaboration. Competitive elaboration is based on the notion that each of the mechanisms has different implications, which can be thought of as indicators or markers that provide evidence about whether the mechanism is working. For example, one mechanism may involve a particular mediator, that is, an intermediate change that should occur in the middle of the causal sequence between program activities and outcomes. For instance, one mechanism (see item 3, described earlier) indicates that kindergarten and first-grade teachers' ratings of children's school readiness will mediate the effects of quality pre-K on longer-term outcomes. A mechanism can also indicate a particular pattern of moderation, whereby the effects are larger for some types of children and families than others. For instance, one mechanism (see item 1 in preceding discussion), which involves reducing negative tracking, should result in stronger effects for children with lower cognitive skills at entry to pre-K (who are thus initially more at risk for negative tracking). Each of the mechanisms points to specific predictions of moderators, mediators, outcome measures, and sequences of change. In short, competitive elaboration involves assessing the extent to which the particular pattern of predictions associated with each mechanism actually occurs.

Among those areas to be measured in carrying out this process are (1) pre-K program quality; (2) social and behavioral skills of the children; (3) cognitive skills of the children; (4) readiness for schooling; (5) expectations for the child's educational attainment; (6) educational resources in the home, including parent interactions and educational materials; (7) program quality in school, beginning with kindergarten; and (8) outcome measures, such as attendance, promotions or retentions, and program assignments. We will also measure certain variables that are subject to state regulation or that may facilitate parental choice of programs for their children and might be related to program outcomes (such as teachers' credentials and type of instructional curriculum). In addition to these measures, certain control variables, such as the socioeconomic status of the families, will be collected because of the pervasive evidence of the influence of family background on educational achievement and attainment.

Consistent with the request for proposals, we propose to follow a cohort of pre-K participants and their families for five years. For most of the students, we anticipate that the four-year-olds in the cohort will progress to the third

grade, the point at which statewide tests of academic skills are available. A brief schedule of surveys, classified by respondent group, and a list of observations that are proposed for systematic data collection are shown in Table 1.1.

Although the request for proposal specifies that a cohort of pre-K participants and their families be followed, it does not specify that a control group of non–pre-K students and families be included in the design. Apart from the issue of budget, the construction of a non–pre-K control group is complicated by the fact that the program is universally available to all preschoolers and their families in the state. Thus, the proposed evaluation emphasizes comparisons within the program, across variations in program curriculum, implementation quality, and type of site (as well as the way such factors differentially affect different types of children). Nevertheless, we propose two strategies for assessing, to the extent possible, the overall effect of the program. First, a small cohort control group will be constructed based on the reports of older siblings of participating children. Second, time series data will be collected on academic outcomes in the early grades, at those school districts at which the records are available, to look for overall improvements related to the level of participation in the pre-K program.

In addition to examining the relationships predicted by those mechanisms identified a priori, we will adopt an additional analytical strategy of principled discovery. Prior research does not provide an adequate base for predictions about how the mechanisms described previously may operate differently for different groups of students. The effect of the program on different subgroups is, of course, an important question. For example, at least some stakeholders would respond differently if the program had a large positive effect on children from disadvantaged households and little effect on children from advantaged households, relative to how they would react if the program had a large positive effect on the advantaged and no effect on the disadvantaged. And differential effects seem plausible. For example, pre-K may increase parental expectations more for students from disadvantaged households than for other children, which could in turn result in more beneficial long-term effects. Methods such as structural equation models can be used to test these differences. Findings about such differences would lead to other sets of predictions, for example, about the change that should occur in parental behaviors across different subgroups of parents. Testing these "second-generation" predictions helps to demonstrate that the original findings (in this example, about parental expectations) were not spurious; hence the term *principled discovery*.

Reporting. We propose several deliverables. First, annual reports will summarize project activities and findings to date, with detailed documentation of the methods and findings. Actual data, without any identifying information, will be made available to the evaluation sponsors and, upon request, to others. A four-page (maximum) brief will also be prepared for policy audiences. During the second year of the project, a parent's guide for choosing a pre-K site for their child will be developed; this will subsequently be updated, as needed,

Table 1.1 Pre-Kindergarten Evaluation: A Pre-Proposal Data Collection Plan

	Year of Study				
Type of Data Collection	*1 Pre-K*	*2 Kindergarten*	*3 1st Grade*	*4 2nd Grade*	*5 3rd Grade*
Classroom Observation	❖	❖			
Program Director/Principal Interviews	❖				
Principal Questionnaire	❖	❖			
Teacher Personal Interviews	❖	❖			
Teacher Telephone Survey	❖	❖	❖	❖	❖
Teacher Administered Tests		❖			
Standardized Tests					❖
Student Progress/Attendance	❖	❖	❖	❖	❖
Parental Telephone Interviews	❖	❖	❖		❖

based on subsequent research findings. In addition, a parent's guide to interacting with the pre-K staff and enhancing their child's development will be developed, using the evaluation findings and other research. Finally, the media will be provided with press releases and reports to convey the findings to the public. It is important for the public and parents to have the best available information to know both the extent to which the program is fulfilling their expectations and the steps taken by the program administrators to enhance the prospect of meeting public expectations. Therefore, the evaluators will work closely with the program administrators in sharing the information with the public.

On Generalizing. For the findings of this evaluation to be useful, they should support generalizations that apply beyond the observed sample to other preschoolers, parents, and preschool sites in Georgia. In addition, Georgia is the first state in the nation to offer a universal, developmental program for four-year-olds, and the results of the proposed evaluation may accordingly be of considerable interest outside as well within the state. In this section, we consider how the results obtained in the proposed evaluation may be useful beyond the specific program sites that are included in the evaluation.

First, we plan to select a probability sample of 200 to 220 of the sites that are currently providing pre-K services to four-year-olds. The sample will be stratified by (1) region of the state; (2) type of organization offering the services (private or local school system); and (3) type of curriculum offered at the site. While the stratification will be complex, it offers several advantages. The sampling strategy will enhance the following factors: the generalizability of the findings within the state; the ability to test for the moderating effects of the stratifying variables; the power of the statistical tests (by way of the lower standard errors, given the sample size); and the credibility of the evaluation. However, while it facilitates generalization within the state of Georgia, the sampling strategy will not in a formal sampling sense allow extrapolating results beyond the state borders.

Extrapolating the results beyond the scope of the sampling frame (in this case, to other states) has noteworthy historical precedent in the area of preschool programs. In particular, many unconstrained generalizations have been based on the results of the Perry Preschool Project in Ypsilanti, Michigan (see Wilson, 1994, for a brief critique). In a very real way, the proposed evaluation contains the basis for generalization as well as for the specification of limits on generalization. Analyses will be provided for different types of children and families (for example, economically advantaged and disadvantaged), settings (such as urban and rural), and curricula. Other states considering a similar pre-K program can see whether the program was successful in those cases similar to the cases that predominate in their populations. In addition, attempts to generalize should be enhanced by the proposed evaluation strategy of linking site-level program activity with the mechanisms that may result in desired outcomes. By improving knowledge on *why* pre-K works as it does, evaluation results should contribute to decisions by other states about whether

to implement a similar program, as well as to decisions within Georgia on how to improve the program.

At the end of an e-mailed set of comments and revisions to the draft proposal, your colleague writes, "I think we're on to something with this emergent realist evaluation. Where does it go from here?" You e-mail back that her question has led you to think a bit about three points related to the role of ERE in the profession of evaluation.

First, according to ERE, the theory and practice of evaluation is in a state of ongoing development (or emergence). Consequently, ERE will appear somewhat different in another decade and will make somewhat different recommendations for practice, in light of forthcoming developments in theory, method, and practice-based lessons. Second, while ER evaluators may sometimes seem evangelical, attempting to "spread the faith" *(as you did with your colleague),* they do not believe in a single, universal religion of evaluation. To the contrary, ER evaluators are happy to see other evaluators following different approaches (such as Patton-style evaluation, which overlaps at least some of the time with organizational development, or Scriven-like assessments of merit and worth without the attention to underlying mechanisms), even though, from the lens of ERE, these types of evaluations have limited usefulness. Still, ERE advocates believe that programs can fruitfully be examined at different levels and at different stages of development, and they see these other approaches as providing aspects of a comprehensive ER evaluation. In this sense, the ER approach to knowledge and values provides a framework for considering the contribution of various approaches to evaluation, and of individual evaluations, in terms of the limitations of the sensemaking technologies employed and the values considered. Use of ERE for such a framework should contribute to that "disputatious community of scholars" (Campbell, 1984) who can talk to each other and debate matters of substance without allowing approach-based differences in terminology to segment conversations into a Tower of Babel. While a healthy profession of evaluation need not be monolithic, forums and language for discussions are vital. And third, as an ER evaluator, you hope that training in ER-based methods, such as methods for values-probing and principled discovery, will become more common.

You sign off your e-mail, smiling to yourself as an old Temptations song plays on the radio. You feel eager to read more about ERE and to apply it to your future evaluation work.

References

Alexander, K., and Entwisle, D. "Educational Tracking During the Early Years: First-Grade Placements and Middle-School Constraints. "In A. Kerckhoff (ed.), *Generating Social Stratification: Toward a New Research Agenda.* Boulder, Colo. Westview Press, 1996.

Barnett, W. "Long-Term Effects of Early Childhood Programs on Cognitive and School Outcomes." *The Future of Children,* 1995, 5 (3), 25–49.

Bhaskar, R. A. *A Realist Theory of Science.* Atlantic Highlands, N.J.: Humanities Press, 1978.

Borkovec, T. D., and Miranda, J. "Between-Group Psychotherapy Outcome Research and

Basic Science." *Psychotherapy and Rehabilitation Research Bulletin,* 1996, 5, 14–20.
Bryk, A. S., and Raudenbush, S. W. "On Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretation." *Psychological Bulletin,* 1988, *104,* 396–404.
Bryk, A. S., and Raudenbush, S. W. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, Calif.: Sage, 1992.
Campbell, D. T. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin,* 1957, *54,* 456–453.
Campbell, D. T. "Pattern Matching as an Essential in Distal Knowing." In K. R. Hammond (ed.), *The Psychology of Egon Brunswik.* New York: Holt, Rinehart and Winston, 1966.
Campbell, D. T. "Qualitative Knowing in Action Research." Kurt Lewin Address; Annual meeting of the American Psychological Association, New Orleans, La., 1974.
Campbell, D. T. "Can We Be Scientific in Applied Social Science?" In R. F. Connor, D. G. Altman, and C. Jackson (eds.), *Evaluation Studies Review Annual,* Vol. 9. Newbury Park, Calif.: Sage, 1984.
Caracelli, V. J., and Greene, J. C. "Crafting Mixed-Method Evaluation Designs." In J. C. Greene and D. J. Caracelli (eds.), *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms.* New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass, 1997.
Carey, T. S., and others. "The Outcomes and Costs of Care for Acute Back Pain Among Patients Seen by Primary Care Practitioners, Chiropractors, and Orthopedic Surgeons: The North Carolina Back Pain Project." *New England Journal of Medicine,* 1995, *333* (14), 913–917.
Carnegie Task Force on Meeting the Needs of Young Children. *Starting Points: Meeting the Needs of Our Youngest Children.* 1994.
Consortium for Longitudinal Studies (ed.). *As the Twig is Bent . . . : Lasting Effects of Preschool Programs.* Hillsdale, N.J.: Erlbaum, 1983.
Cook, T. D. "Postpositivist Critical Multiplism." In R. L. Shotland and M. M. Mark (eds.), *Social Science and Social Policy.* Beverly Hills, Calif.: Sage, 1985.
Cook, T. D. "A Quasi-Sampling Theory of the Generalization of Causal Relationships." In L. B. Sechrest and A. G. Scott (eds.), *Understanding Causes and Generalizing About Them.* New Directions for Program Evaluation, no. 57, 1993.
Cook, T. D., and others. *Meta-Analysis for Explanation: A Casebook.* New York: Russell Sage Foundation, 1992.
Cook, T. D., and Campbell, D. T. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally, 1979.
Cronbach, L. J. "Beyond the Two Disciplines of Scientific Psychology." *American Psychologist,* 1975, *30,* 116–127.
Cronbach, L. J. *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass, 1982.
Entwisle, D. "The Role of Schools in Sustaining Early Childhood Program Benefits." *The Future of Children: Long-Term Outcomes of Early Childhood Programs,* 1995, 5 (3), 133–144.
Gamoron, A. "Educational Stratification and Individual Careers." In A. Kerckhoff (ed.), *Generating Social Stratification: Toward a New Research Agenda.* Boulder, Colo.: Westview Press, 1996.
Greene, J. C., and Caracelli, V. J. "Defining and Describing the Paradigm Issue in Mixed-Method Evaluation." In J. C. Greene and D. J. Caracelli (eds.) *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms.* New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass, 1997.
Guba, E .G., and Lincoln, Y. S. *Fourth-Generation Evaluation.* Newbury Park, Calif.: Sage, 1989.
Hanushek, 1997
Hendricks, M. "Making a Splash: Reporting Evaluation Results Effectively." In J. S. Wholey, H. P. Hatry, K. E. Newcomer (eds.) *Handbook of Practical Program Evaluation.* San Fransisco, Jossey-Bass, 1994.

Henry, G. T. *Graphing Data: Techniques for Display and Analysis.* Thousand Oaks, Calif.: Sage, 1995.

Henry, G. T. "Community-Based Accountability: A Theory of Accountability and School Improvement." *Phi Delta Kappan,* 1996, *78* (1), 85–90.

Henry, G. T. *Creating Effective Graphs: Solutions for a Variety of Evaluation Data.* New Directions for Evaluation, no. 73. San Francisco: Jossey-Bass, 1997.

Henry, G., and Bugler, D. *Evaluation of the Georgia HOPE Scholarship Program: Impact on Students Attending Public Colleges and Universities.* Atlanta, Ga.: Council for School Performance Report, 1997.

Henry, G. T. and others. *School Performance Reports.* Atlanta, Ga.: Council for School Performance. 1996.

House, E. R. *Jesse Jackson and the Politics of Charisma: The Rise and Fall of the PUSH/Excel Program.* Boulder, Colo.: Westview, 1988.

House, E. R. "Realism in Research." *Educational Researcher,* 1991, *20,* 2–9.

Jones, J., Dolan, K., and Henry, G. *Two Miles Down a Ten-Mile Road: Instructional Technology and the Impact of Lottery Funding in Georgia.* Council for School Performance, 1996.

Julnes, G. "Context-Confirmatory Methods for Supporting Disciplined Induction in Post-Positivist Inquiry." Paper presented at the annual meeting of the American Evaluation Association, Vancouver, British Columbia, November 2, 1995.

Kingdon, J. *Agendas, Alternatives, and Public Policies.* (2nd ed.) HarperCollins, 1995.

Lee, V., and Leob, S. "Where Do Head Start Attendees End Up? One Reason Why Preschool Effects Fade Out." *Educational Evaluation and Policy Analysis,* 1995, *17* (1), 62–82.

Majone, G. *Evidence, Argument, and Persuasion in the Policy Process.* New Haven, Conn.: Yale University Press, 1988.

Marcon, R. "Differential Effects of Three Preschool Models on Inner-City Four-Year-Olds." *Early Childhood Research Quarterly,* 1992, *7,* 517–530.

Marcon, R. *Early Learning and Early Identification Follow-Up Study: Transition from the Early to the Later Childhood Grades.* Washington, D.C., 1994, 1–22. (ED 263984)

Mark, M. M. "Validity Typologies and the Logic and Practice of Quasi-Experimentation." In W.M.K. Trochim (ed.), *Advances in Quasi-Experimental Design and Analysis.* New Directions for Program Evaluation, no. 31. San Francisco: Jossey-Bass, 1986.

Mark, M. M. "From Program Theory to Tests of Program Theory." In L. Bickman (ed.), *Advances in Program Theory.* New Directions for Program Evaluation, no. 47. San Francisco: Jossey Bass, 1990.

Mark, M. M. "On the Integration of Discovery, Confirmation, and Monitoring in Evaluation." Paper presented at the Trinity Symposium on Public Management Research. San Antonio, Tex., 1995.

Mark, M. M., Feller, I., and Button, S. B. "Integrating Qualitative Methods in a Predominantly Quantitative Evaluation: A Case Study and Some Reflections." In J. C. Greene and D. J. Caracelli (eds.), *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms.* New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass, 1997.

Mark, M. M., Henry, G. T., and Julnes, G. "Evaluation, A Realist Approach: Monitoring, Classification, Causal Analysis, and Values Inquiry." In preparation.

Mark, M. M., Hofmann, D., and Reichardt, C. S. "Testing Theories in Theory-Driven Evaluations: (Tests of) Moderation in All Things." In H.-t. Chen and P. H. Rossi (eds.), *Using Theory to Improve Program and Policy Evaluations.* New York: Greenwood Press, 1992.

Maxwell, J. *Qualitative Research Design: An Interactive Approach.* Thousand Oaks, Calif.: Sage, 1996.

McKey, R. *The Impact of Head Start on Children, Families, and Communities.* Washington, D.C., 1985. (ED 263984)

Mohr, L. B. *Impact Analysis of Program Evaluation.* (2nd ed.) Thousand Oaks, Calif.: Sage, 1995.

Patton, M. Q. "Paradigms and Pragmatism." In D. M. Fetterman (ed.), *Qualitative Approaches to Evaluation in Education: The Silent Scientific Revolution.* New York: Praeger, 1988.

Patton, M. Q. *Utilization-Focused Evaluation: The New Century Text.* Thousand Oaks, Calif.: Sage, 1997.

Pawson, R., and Tilley, N. *Realistic Evaluation.* Thousand Oaks, Calif.: Sage, 1997.

Reichardt, C. S., and Mark, M. M. "Quasi-Experimentation." In L. Bickman and D. J. Rog (eds.), *Handbook of Applied Social Research Methods.* Thousand Oaks, Calif.: Sage, 1998.

Ross, L. and Nisbett, R. E. *The Person and the Situation.* New York: McGraw Hill, 1991.

Ross, H. L., Campbell, D. T., and Glass, G. V. "Determining the Social Effects of Legal Reform: The British 'Breathalyser' Crackdown in 1967." *American Behavioral Scientist,* 1970, *13,* 493–509.

Scriven, M. *The Evaluation Thesaurus.* Thousand Oaks, Calif.: Sage, 1990.

Sechrest, L., and Yeaton, W. E. "Assessing the Effectiveness of Social Programs: Methodological and Conceptual Issues." In S. Ball (ed.), *Assessing and Interpreting Outcomes.* New Directions for Program Evaluation, no. 9. San Francisco: Jossey Bass, 1981.

Shadish, W. R., Cook, T. D., and Leviton, L. C. *Foundations of Program Evaluation: Theories of Practice.* Newbury Park, Calif.: Sage, 1991.

Smith, M. L. "Mixing and Matching: Methods and Models." In J. C. Greene and D. J. Caracelli (eds.), *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms.* New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass, 1997.

Smith, N. L. "Aspects of Investigative Inquiry in Evaluation." In N.L. Smith (ed), *Varieties of Investigate Evaluation.* New Directions for Program Evaluation, no. 50. San Francisco: Jossey Bass, 1992.

Trochim, W.M.K. "Pattern Matching, Construct Validity, and Conceptualization in Program Evaluation." *Evaluation Review,* 1985, *9,* 575–604.

Tukey, J. W. *Exploratory Data Analysis.* Reading, Mass.: Addison-Wesley, 1977.

White, L. "Policy Analysis as Discourse." *Journal of Policy Analysis and Management,* 1994, *12* (3), 322–359.

Wholey, J. S. "Evaluability Assessment: Developing Program Theory." In L. Bickman (ed.), *Using Program Theory in Evaluation.* New Directions in Program Evaluation, no. 33. San Francisco: Jossey-Bass, 1987.

Wilson, J. "Culture, Incentives, and the Underclass." In H. J. Aaron, T. E. Mann, and T. Taylor (eds). *Values and Public Policy.* Washington, D.C.: The Brookings Institution, 1994.

Wright, J. D. "Methodological Issues in Evaluating the National Health Care for the Homeless Program." In D. J. Rog (ed.), New Directions for Program Evaluation, no. 52. San Francisco: Jossey-Bass, 1991.

Yankelovich, D. "How Changes in the Economy are Reshaping American Values." In H. J. Aaron, T. E. Mann, and T. Taylor (eds), *Values and Public Policy.* Washington, D.C.: The Brookings Institution, 1994.

Yin, R. K. *Case Study Research: Design and Methods.* (2nd ed.) Thousand Oaks, Calif.: Sage, 1994.