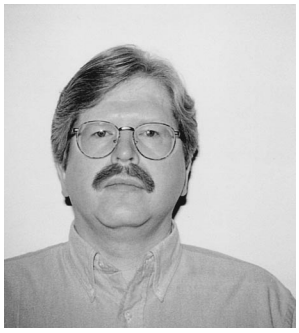
Articles should deal with topics applicable to the broad field of program evaluation. Articles may deal with evaluation practice or theory, but if the latter, implications for practicing evaluators should be clearly identified. Examples of contributions include, but are not limited to, reviews of new developments in evaluation and descriptions of a current evaluation effort, problem, or technique. Results of empirical evaluation studies will be published only if the methods and/or findings have clear utility to other evaluation practitioners. Manuscripts should include appropriate references and normally should not exceed 10 double-spaced typewritten pages in length; longer articles will occasionally be published, but only where their importance to AJE readers is judged to be quite high.

Toward an Integrative Framework for Evaluation Practice[1]

MELVIN M. MARK, GARY T. HENRY, AND GEORGE JULNES



Melvin M. Mark

ABSTRACT

Evaluation has been beset with serious divisions, including the paradigm wars and the seeming segmentation of evaluation practice into distinct evaluation theories and approaches. In this paper, we describe key aspects of an integrative framework that may help evaluators move beyond such divisions. We offer a new scheme for categorizing evaluation methods within four *inquiry modes*, which are "families" or clusters of methods: description, classification, causal analysis, and values inquiry. In addition, we briefly describe a set of alternative *evaluation purposes.* We argue that, together with a form of realist philosophy, the framework of inquiry modes and evaluation purposes (1) provides a common lexicon for evaluators, which may help the field in moving beyond past divisions, and (2) offers a useful approach to evaluation planning.

The field of evaluation has no shortage of distinctions and divisions. Most notable among these is the schism that has even been described commonly in the metaphoric language of

Melvin M. Mark • Institute for Policy Research & Evaluation, N–254 Burrowes, Pennsylvania State University, University Park, PA 16802; E-Mail: M5M@psu.edu; **Gary T. Henry** • School of Policy Studies, Georgia State University, University Plaza, Atlanta, GA 30303. Email: gthenry@gsu.edu; **George Julnes** • Department of Psychology, Utah State University, Logan, UT 84322-2810.

American Journal of Evaluation, Vol. 20, No. 2, 1999, pp. 177–198.
ISSN: 1098-2140

battle: the "paradigm war." The "sides" in the paradigm war are most often described in terms of methodology, that is, qualitative versus quantitative. At other times, the distinction is made at more of a philosophical level, with constructivists and postmodernists highlighting their differences with the sons and daughters of positivism, and vice versa.

In the last two decades, much of the energy and intellectual capital in the field of evaluation has been spent in skirmishes and in peace talks between advocates of qualitative and quantitative paradigms. The paradigm war has been the subject of numerous plenary sessions, presidential addresses, papers, chapters, volumes, and hallway conversations. Sometimes the combatants in this war have argued that there can be no accommodation between the sides. For example, Lincoln (1990) said that "accommodation between paradigms is impossible. The rules for action, for process, for discourse, for what is considered knowledge and truth, are so vastly different that, although procedurally we may appear to be undertaking the same search, in fact, we are led to vastly diverse, disparate, distinctive, and totally antithetical ends" (p. 81). One might surmise that the redundancy (and repetitiveness) in the phrase "vastly diverse, disparate, distinctive, and totally antithetical" reflects the intensity with which this view was held.

Despite Lincoln's (1990) and other claims of incompatibility, there is also a long history of calls for integration across the two paradigms (Reichardt & Cook, 1979). And of late, there is more talk of peace than of war (e.g., Caracelli & Greene, 1997; Datta, 1994; House, 1994; Mark, Feller, & Button, 1997; Reichardt & Rallis, 1994; M. L. Smith, 1994, 1997). Even some who have argued that accommodation is impossible have subsequently acknowledged the possibility of reconciling the qualitative and quantitative paradigms (Lincoln & Guba, 1994, p. 189). But all may not yet be well. The rhetorical artillery may have stopped firing, but it is not clear that the groundwork has been laid for a lasting peace. As some of the most insightful commentators about the war have noted, an important question remains: Under what conceptual stand or paradigm is the integration to occur (Datta, 1997; Smith, 1997)?

This question is all the more difficult to answer because, while the paradigm schism has pre-occupied many in the field, the theory and practice of evaluation has become quite Balkanized (in the sense of an increasing number of boundaries dividing small groups that may be in conflict). There are numerous "named" evaluation theories and approaches (theory-driven, responsive, summative, empowerment, goal-free, C.I.P.P., emancipatory, investigatory, outcome, fourth-generation, participatory, utilization-focused, et al.), most of which are claimed as somehow distinctive by their advocates. The number of supposedly distinct approaches is sufficiently large that it may surpass the number of countries that have in fact ever existed in the Balkan region. The large and growing list of evaluation theories (or approaches) can give the impression of an uneasily fragmented field, with many boundaries separating different cultures and practices.

There are of course positive interpretations that can be made of the large and increasing number of evaluation theories and approaches. This can be a sign of a field in a period of exciting intellectual development. New approaches may be arising in response to emerging or previously ignored needs of service recipients, evaluation clients, and other stakeholders. In addition, a multiplicity of evaluation theories can provide flexible evaluators with a rich menu of approaches from which one can be selected that fits well with the needs and demands of a particular evaluation.

Despite these and other positive interpretations, the image and reality of a Balkanized field presents a negative side. With a seemingly unlimited number of approaches to evaluation, prospective clients (and even evaluators) may have trouble knowing what their choices

are, or even what evaluation is. Balkanization also often begets skirmishes and tribal warfare. Reasoned debate is often difficult, because those in different camps may speak different languages, so that words are exchanged but ideas are not.

Under these circumstances, moreover, there are few widely-shared models of how to make reasoned choices when planning an evaluation design. Evaluators from different camps may bring vastly different views of what type of evaluation should be done in a given situation. The evaluator's background may be the most important determinant of the type of evaluation that is done, rather than the context and the information needs of affected groups and the public. Although we will address this issue only tangentially in this paper, we believe the broader theory of evaluation sketched here provides a useful framework for planning evaluations, including approaches suggested by many diverse evaluation theories. One benefit of an overarching framework of this sort is that the strengths, weaknesses, and trade-offs involved in choosing one approach or the other are more explicit and can better be communicated to all involved.

In this article, we sketch selected aspects of a larger framework that, we believe, can serve as a foundation for integration. We offer a new scheme for categorizing evaluation methods, placing them into four different *inquiry modes*, which are "families" or clusters of method, each with a different general objective. The framework of inquiry modes provides part of the language needed to move beyond the qualitative-quantitative dichotomy. In addition, we suggest that a framework of *evaluation purposes* is another needed part of a common lexicon for evaluators. Evaluation purpose, as is briefly discussed, also serves as one of the contextual features that should guide the choice of inquiry mode. We further suggest (in a very abbreviated form here) that the philosophy of science known as realism can provide a common paradigm for the combatants in the paradigm war. With a common language for discourse that transcends past wars and divided camps, one can then plan for individual evaluations and evaluation programs based on the needs *in situ*, rather than on the predispositions of whoever is doing the evaluation.

EVALUATION AS ASSISTED SENSEMAKING

The practice of evaluation, we believe, should be seen as a form of assisted sensemaking. As assisted sensemaking, evaluation is an adjunct to and in support of natural sensemaking. Humans have evolved (or, if you prefer, Creator-endowed) capacities for natural sensemaking. Sensemaking capabilities allow humans to observe regularities, to develop accounts as to why regularities occur, and to undertake behaviors designed to capitalize on that which has been learned. These capacities, however, are limited. In response to these limits, humans have constructed technologies that assist our valuable but imperfect natural sensemaking capabilities. For example, eyeglasses have been created to assist natural vision. Number systems and mathematics have been developed to assist our natural abilities to quantify (for evidence of basic counting even in infants, see Wynn, 1992). Similarly, the field of evaluation has been developed to assist and extend natural human abilities to observe, understand, and make judgments about policies, programs, and other objects of evaluation.

Sensemaking has (at least) two components, representational and valuative, that appear to occur naturally in humans. The *representational component of sensemaking* refers to human efforts to know and understand what is going on in the world around them. Humans observe events, categorize, and make causal judgments. They observe patterns, develop

hypotheses, and move back and forth between evidence and explanations. The *valuative component* refers to human tendencies to make judgments about what is better and what is worse. People have their own subjective experiences and have empathy for others. They like some things more than others, make moral judgments, weight values, and even consider what might be best for the common good.

The people who establish, administer, or are affected by policies and programs, without any special urging, engage in natural sensemaking about them. However, individuals are limited in their capacity to do so. They often cannot, for example, observe all program sites. Rarely could an individual find, observe, and record, for all consumers of services, the outcomes that followed years after they were program clients. Furthermore, individual, unsystematic observations are likely to be biased in any number of ways. For example, people's expectations and desires often bias the way they notice, interpret, and recall information. The four inquiry modes (described hereafter) have evolved as assisted sensemaking techniques that enable humans to compensate for the limits of their own natural sensemaking. While all assisted techniques are fallible and have inherent flaws, they have been developed to overcome many of the limitations of natural sensemaking.

We do not claim to have the last word on sensemaking and its components. Finer distinctions could certainly be drawn, distinguishing, for example, between various attentional, perceptual, memory, and higher-order cognitive systems that operate on the representational side of sensemaking. In addition, ongoing developments in psychology and elsewhere might lead to a different formulation of sensemaking in the future. Nevertheless, the general view offered here is, we believe, useful in the view it provides of evaluation as assisted sensemaking, in service both of the representational and evaluative components of natural sensemaking.

FOUR MODES OF INQUIRY

Inquiry modes are distinguishable groupings of methods used by evaluators (and others). They are clusters or families of methods, each of which produces a different type of information. Each inquiry mode represents a set of methods—sensemaking techniques—that have evolved in evaluation and in other types of systematic social inquiry. Inquiry modes do not, however, correspond to the traditional qualitative-quantitative distinction. To the contrary, each of the four inquiry modes includes both quantitative and qualitative methods within it.

We argue that the methods that most evaluators have traditionally carried out under the umbrella of program and policy evaluation fall within three distinct inquiry modes: description, classification, and causal analysis. These three inquiry modes assist the representational component of natural sensemaking. In addition, our approach calls for the inclusion of another mode of inquiry that has less often been used in many past approaches to evaluation practice: values inquiry. Values inquiry, which assists the valuative component of sensemaking, is the systematic, usually empirical study of the values embedded within social programs.

We discuss each of four inquiry modes in somewhat more detail in the remainder of this section. Although we shall, initially at least, describe these as though they are completely distinct sets of methods, as we note briefly later, things are somewhat more complicated. Before addressing the ways in which they can overlap, unintentionally or by design, let us

consider the nature of each of the four inquiry modes, along with illustrations of the methods that belong within each.

Description

Description refers generally to a cluster of methods that can assess the status of some observable attribute. Description occurs when events and experiences are observed, recorded, or otherwise measured, and when these data are organized to describe patterns and discrepancies from patterns. Qualitative evaluators have long emphasized the description of settings as a way of representing complexity (see, e.g., Spradley, 1980). The goal of much qualitative observation is to develop a rich description of the program, its context, and how it is viewed by its clients. Quantitative researchers, in contrast, often seem to judge description as less valuable than other inquiry modes. However, performance measurement and monitoring have recently elevated the role of description among quantitatively-oriented evaluators. Performance measurement and monitoring systems measure selected aspects of (1) program clientele (the number of clients, and client characteristics such as demographics), (2) services delivered (i.e., the amount, type, and means by which services are delivered), (3) program resources and (4), increasingly, program outcomes. With respect to outcomes, performance measurement systems usually do not make comparisons that allow the observed outcomes to be attributed confidently to the program (unlike in causal analysis, where this concern is central).

Perhaps because of its limitations in the attribution of cause, performance measurement has not been given significant attention in evaluation theory (with the important exception of the work of Wholey and his colleagues, e.g., Wholey, 1983). However, movements such as the Government Performance and Results Act (GPRA), state-level performance measurement requirements (Melkers & Willoughby, 1998), and not-for-profit agencies' adoption of performance measures have brought a growing recognition of the importance of this type of description in evaluation.

A wide array of methods can be brought to bear for description, including direct observation, structured and unstructured interviews, focus groups, and surveys using probability samples. Often management information systems or other administrative data provide information for description. Sometimes these allow evaluators to assess trends and make comparisons across groups or program sites. In many content areas, such as childcare and preschool programs (Scarr, Eisenberg, & Deater-Deckard, 1994), highly structured, on-site observation techniques are used to record and measure actual service delivery.

Description includes one-time observations of programs as well as ongoing, periodic monitoring. For example, "performance audits" are one-time measurements of outcome variables, as when the U.S. General Accounting Office (GAO) assesses, on a one-time basis, the outcome variables associated with some program. Needs assessment also relies on the descriptive mode of inquiry.

Classification

Description focuses on observable, measurable attributes of individuals or social aggregates. Classification is intended to tap deeper structures. Classification refers, generally, to methods used to assess underlying structures and categories, and for techniques for determining to which category an individual belongs. In the context of evaluation, classifi-

cation is used (1) to identify the categories into which clients, services, programs, times, settings or effects fall, and (2) to assign specific cases (of services, or programs, or individuals, or settings, or outcomes) to general categories or groupings. The categories themselves cannot be directly observed or measured. Instead, they must be inferred from the patterns of observable characteristics.

In the context of evaluation, a number of important tasks involve classification, as illustrated by the following questions: Does the Bank Street preschool program use child-initiated or adult-directed teaching models (Marcon, 1992, 1994). Is a particular local preschool site truly an example of the Bank Street curriculum? What are the components (Wholey, Hatry, & Newcomer, 1994) or elements (Cook & Shadish, 1986) that make up a given program?

Qualitative, quantitative, and mixed methods can all be used in classification. In evaluation (as elsewhere), classification involves two steps. First, the different categories must be inferred, along with the description-level characteristics that are used to define a category. Methods such as case studies, exploratory and confirmatory factor analysis, and cluster analysis can be used to determine the classification system, which may also be referred to as a taxonomy or category system. For example, Marcon (1992, 1994) has developed a set of 14 items for classifying pre-school teachers into one of three models or categories: child-initiated, adult-directed, and middle-of-the-road. As another example, Orwin, Sonnefeld, Cordray, Pion, & Perl (1998) used an expert panel, referred to as the Taxonomy Panel, to generate a classification of 39 alcohol and drug abuse services for homeless people.

In a second possible step of classification, the evaluator may determine whether a specific case falls within one class. This is done, generally, by comparing the attributes of the specific case with the pattern of attributes used to define the category. This second classification step includes questions about (1) whether a given program falls within some broader, conceptual category; (2) whether a local project falls within the established bounds of a pre-existing program or curriculum; and (3) whether one set of local projects falls within a different program sub-type than do other projects. For example, Georgia Pre-K programs have been classified into three categories using Marcon's methods (Henderson, Basile, & Henry, 1999). This is part of a larger, longitudinal evaluation that will examine the patterns of outcomes associated with each of the three teaching models.

Causal Analysis

Causal analysis refers to a variety of methods used to investigate the causal relationships set into play by a program or policy. Traditionally, causal analysis in evaluation has primarily involved estimating the impact of a program on selected outcomes. For example, one might assess whether the Bank Street preschool program has positive effects, including cognitive gains. Assessment of effects goes beyond simple monitoring of outcomes, in that alternative interpretations of the outcomes, other than the program (e.g., differences in family background, other historical events), must be ruled out so that the outcome can be attributed to the program. In many case study evaluations, this attribution of causation is supported through an iterative pattern-matching process that develops explanations, deduces other implications of those explanations, and then seeks additional information that addresses these deduced implications (Yin, 1994). In evaluations that make greater use of quantitative data, causal analysis—unlike monitoring—generally requires the collection of data from compar-

ison groups along with other procedures (e.g., controlled assignment of individuals to groups) to support the claim that the program, and not something else, is responsible for the outcomes observed.

Another important aspect of causal analysis is the identification of the mechanisms that underlie any effects. For example, if a public preschool program has positive effects on children's readiness for first grade and their staying on grade level, we would want to know why. Was it because the program raised teachers' expectations, or because of specific skills the children developed, such as communication skills, or because of some other underlying mechanism (Henderson et al., 1999)? If we can explain why effects occur, we should be better able to generalize and to use the findings in other settings (Cook, 1993; Cronbach, 1982; Mark, 1990).

Values Inquiry

Description, classification, and causal analysis support natural representational sensemaking. The fourth inquiry mode, in contrast, assists the valuative side of natural sensemaking. Values inquiry refers to a variety of methods that can be applied to the systematic assessment of the values positions surrounding the existence, activities, and outcomes of a social policy or program. Natural valuation capacities rely on direct experience or empathy. Evolved technologies have been developed for unearthing values and analyzing the assumptions and limitations of various value positions. Values inquiry can be useful, not only in supporting natural valuation, but also in the design and interpretation of evaluation activities using other inquiry modes. For example, the results of values inquiry can be used to prioritize the effects that an evaluation will address, point to the limitations of the effects addressed in prior evaluations, or guide the reporting of various evaluation findings. Our view is that values, especially systematic inquiry about values, have been given inadequate attention in most previous evaluation writings, at least by those who emphasize such traditional sensemaking activities as description, classification, and causal analysis.

Usually, values inquiry will involve measurement of the extent to which various stakeholder groups, and the public, value various possible outcomes or attributes of the program or policy. For example, an evaluation of a preschool program might examine the extent to which parents, program staff, and the public value such possible outcomes as improved test scores, social skill development, enhanced family interactions, reduced school drop-out rates, reduced delinquency, and decreases in the gap between advantaged and disadvantaged children's school performance. In some circumstances, surveys could be used to obtain ratings and rankings of the potential outcomes (Henry, 1999). Surveys, for example through telephone interviews, also allow analyses of how values differ across region, education, occupation, or other possible factors of interest. Qualitative approaches to values inquiry encompass a wide range of methods. These range from group interviews and focus groups with members of different stakeholder groups, who may be asked to deliberate about the possible outcomes and how they might be weighed and rated, to analysis of mission statements to reveal embedded values (Flick, 1998).

We have listed three inquiry modes on the representational side of sensemaking, and only one on the valuative side. This is *not* intended to indicate that the valuative part of sensemaking is less important. To a large extent, this imbalance simply reflects the greater attention historically to the development of systematic methods to assist representational sensemaking. In fact, by including values inquiry as an inquiry mode, we intend to emphasize

TABLE 1.
Summary of the Four Inquiry Modes

	Description	*Classification*	*Causal Analysis*	*Values Inquiry*
Focus	(ongoing) measurement of clients, processes, and outcomes	putting case (e.g., projects) into categories (e.g., program subtypes)	estimate effects of program; probe mechanisms	identify values positions of stakeholders and public
Typical quantitative method	management information system	cluster/factor analysis	experiment	survey
Typical qualitative method	observation	comparative case study	iterative pattern matching	focus groups with stakeholders

its importance. We believe that attention to assisting the valuative side of sensemaking is usually at least as important as assisting the representational side. The inclusion of values inquiry as a separate inquiry mode has another benefit. It helps make this framework more useful as an aid to clearer communication across evaluation camps. In the past, confusion between values inquiry and other inquiry modes has, we believe, contributed to many occasions in which evaluators from different traditions simply talked past each other.

Table 1 provides, for each the four inquiry modes, the focus of the inquiry mode, along with a typical qualitative and a typical quantitative research method that falls within that mode. The four inquiry modes are distinct but not mutually exclusive, as we discuss shortly. Nevertheless, delineating the inquiry modes, particularly values inquiry, help us to understand the proper role of evaluation as assisted sensemaking.

Values Inquiry and the Proper Role of Evaluation

We earlier described evaluation as a form of assisted sensemaking. An implication of this perspective is that evaluation should serve as an aid to others' judgments, not as a substitute for them. In the context of public programs and policies, the relevant judgments are carried out through democratic processes and institutions, including elected representatives, appointed officials, public debate, and elections (for more detailed discussion, see Mark, Henry, & Julnes, in press). The inquiry modes match well with the sort of questions that parties in deliberations about programs and policies are likely to have, including: (1) what services are delivered to whom (description); (2) what if any different types of services are being offered (classification); (3) what if any effects do the services have, and why (causal analysis); and (4) who cares most about what issues related to the services (values inquiry)?

Values inquiry in particular can increase the contribution of evaluation to these democratic processes and institutions. We will only briefly illustrate this here. We start with the assumption that those involved in democratic processes, including direct stakeholders, citizens, and public officials, naturally "valuate." Moreover, they have natural abilities to make value judgments, in pursuit of their view of social betterment, whatever that may be. Typically, however, the issues surrounding programs and policies are sufficiently complex

that these natural abilities can benefit from assistance. One form of assistance that values inquiry can provide is to improve information about the value positions held by others. Findings from values inquiry can be an important *result*. These findings will often warrant presentation to the various evaluation audiences. This is not to say that findings about evaluation inquiry are a magic bullet. Rather, our claim is simply that the results of values inquiry can support natural valuation.

Values inquiry can contribute in other ways as well. Values inquiry early in an evaluation project can also contribute to planning of subsequent evaluation work using a different inquiry mode. For instance, values inquiry can indicate which outcomes are most important to stakeholders and the public and, in some cases, identify unintended effects.

Another potential contribution of values inquiry is to help shape evaluation reports and other forms of communication that provide the types of information that are most relevant to people's natural valuation. For example, assume that a values inquiry reveals that stakeholders are particularly concerned with certain subgroups in the population of program clients. Stakeholders might, for instance, be particularly concerned with homeless *families*, or be especially interested in the health of the *children* of welfare recipients. An evaluation would then best support natural valuation if it reports both overall results and separate findings for these groups, so that these more important considerations can more easily be part of the valuation process.

Realism, Practice, and the Rationale for the Four Inquiry Modes

The distinction we have discussed between the four inquiry modes developed from our own review of the methods used in evaluation practice. As we began to work with the practice-based four-fold classification, we were delighted to see that our distinctions paralleled ones found in the realist philosophy of science. For reasons summarized later, we believe that realism has much to offer for evaluation, including a possible resolution of the paradigm wars. Accordingly, we take some satisfaction—and believe there is some validation—in the grounding that realist philosophy provides for the four inquiry modes.

Contemporary realists often refer to the existence of a reality that is structured, hierarchical, and embedded (e.g., Bhaskar, 1975, 1998; House, 1991; Pawson & Tilley, 1997). In large part, this means that there are unobservable phenomena that underlie the things we perceive and experience. Beneath the surface, so to speak, there are structures, categories of things (e.g., at one level of analysis, oxygen and other elements). There are also underlying generative mechanisms (such as combustion). And the underlying structures and mechanisms give rise to what we perceive and experience (e.g., fire).

The inquiry mode of classification corresponds directly to the realist notion of structures. That is, methods for classification have been developed to help discover and demonstrate meaningful groupings of objects in our world. The inquiry mode of causal analysis corresponds directly to the realist notion of underlying generative mechanisms. That is, methods for causal analysis have been developed to probe the unobservable causal connections in which we are interested. The inquiry mode of description corresponds roughly to the realist concept of a more directly perceived and experienced level of reality. In the language of some realists (e.g., Bhaskar, 1975), activities at this level are referred to as events or, when observed, experiences. Methods for description focus more on recording, counting and measuring services, clients, and effect variables, with (typically) fewer inferences drawn about underlying latencies or unobservables (such as causation and category structure).

We identify values inquiry as a separate inquiry mode, not because of the nature of the physical world, but because humans are constantly assessing and assigning value to their experiences and to events in the physical realm, as realists such as Putnam (1987, 1990, 1994) emphasize. As noted above, when humans observe events, such as social conditions, they naturally make judgments that some are better than others, that some are more desirable, that some are necessary for the "common" good. Values inquiry extends these natural capacities and tendencies. People's views of social programs, for example, clearly depend on observed events involving the program and on the structures and mechanisms thought to underlie the program. However, they also depend upon the human reactions to the observed and inferred characteristics of the program. In other words, evaluation findings about structures, mechanisms, and events are filtered through a lens of human values. As a result, findings from values inquiry can assist the reasoning done within democratic institutions, as they carry out the processes whereby social betterment is defined and sought.

Because the rationale for values inquiry differs from that of the other inquiry modes, the methods for values inquiry may overlap with the methods from other modes. For example, one can apply methods of description to human values rather than to social programs and policies. Accordingly, values inquiry can involve survey and focus group methodologies that can also be used for description. This is not, however, the only way in which the lines between inquiry modes, which we have presented so far as sharp, can become blurred.

Linkages Across Inquiry Modes

There are several ways in which the inquiry modes are, or can appear to be, interrelated. The first is that the inquiry modes, which consist of techniques developed by humans to assist sensemaking with systematic inquiry, all depend on *natural* sensemaking. Each of the inquiry modes has analogs in natural sensemaking. For example, the systematic methods of classification have a natural analog in categorization—the sorting of objects into categories—which is a natural activity for humans (e.g., Rosch & Bloom, 1978) and underlies human cognition and action. Natural categorization underlies any systematic inquiry aimed at description, causal analysis, or values inquiry. For example, when doing description, evaluators (and others) naturally categorize some people as "program clients." In parallel fashion, other aspects of natural sensemaking correspond to the four inquiry modes. The systematic methods of a given inquiry mode may in fact depend on other kinds of natural sensemaking. Nevertheless, the inquiry modes represent systematic inquiry methods, developed and refined by humans. While natural categorization might underlie other inquiry modes, the assisted sensemaking techniques that have evolved to systematically carry out classification, such as cluster and factor analysis, can usefully be set apart from those of the other inquiry modes. Moreover, the formal inquiry modes have far less overlap than do the corresponding natural tasks of description, categorization, causal attribution, and valuing.

Second, inquiry modes can be combined, either within a single evaluation, or in a sequence of evaluations. For example, an evaluator might take the product of an evaluation that involves description, such as MIS data on the attributes of local projects and their clients, and use that in a cluster analysis to conduct classification (see, e.g., Conrad & Beulow, 1990). As another example, while classification will sometimes be the aim of an entire evaluation, in other cases it will be sequenced with one or more other mode of inquiry. For example, one might classify local projects into categories prior to conducting causal analysis to assess the

effects of each subtype of the program (e.g., Henry, Basile, & Henderson, 1999). Such combinations and sequences also exist for the other modes.

The linkages between values inquiries and the other inquiry modes deserve special attention. When conducted prior to another inquiry mode, values inquiry can help shape the other investigation, as noted above. Values inquiry can help, for example, in selecting the effects that need to be considered in causal analyses. In areas such as preschool, welfare, and negative income tax, the initial evaluations have focused on a restricted set of outcomes (e.g., cognitive gains in the case of preschool); these have subsequently become expanded in later evaluations (e.g., retention at grade level, and involvement with the criminal justice system). While a systematic values inquiry would not guarantee that all effects of interest would be identified initially, deliberative processes and interpretive methods of values inquiry may increase the likelihood that a well-rounded list of important effects is developed before other (usually costly) modes of evaluation begin. In a similar vein, values inquiry can guide description and classification.

In this respect, values inquiry is an elaboration of previous work on stakeholder-based evaluation (e.g., Bryk, 1983; Mark & Shotland, 1985; Weiss, 1983). Stakeholder-based evaluation and more recent variants also highlight the importance of up-front attention to the concerns of stakeholders. However, rather than viewing such input as an aspect of evaluation *process*, as stakeholder approaches generally do, we identify values inquiry as a separate inquiry mode. In doing so, we highlight not only that there are systematic methods for studying values, but also that values inquiry provides results that may be interesting and of value in their own right.

A third way that lines between the inquiry modes can seem blurry is common to most if not all taxonomies (Putnam, 1990). As with other taxonomies, there are individual cases that are near the boundaries between categories. For example, data from the description mode can, under some circumstances, be used for causal analysis. On the quantitative side, data from an ongoing monitoring system can be used in an interrupted time series quasi-experiment. In the qualitative vein, data from the description mode of inquiry are critical in iterative pattern-matching approaches to causal inference. Using inquiry modes together or in sequence creates some fuzzy or hybrid cases, but this does not, we believe, threaten the usefulness of the distinctions.

Inquiry Modes: Beyond Past Divisions in Evaluation

We opened this article with a discussion of the paradigm wars between the "quals" and the "quants." The framework of inquiry modes can, we believe, contribute to a lasting peace. In no small part, this is because there are both qualitative and quantitative (and mixed) approaches within each of the inquiry modes. And recognition of common goals is one of the classic methods of moving beyond intergroup conflict and creating instead a sense of cohesion (Sherif, Harvey, White, Bond, & Sherif, 1961).

In addition, the framework of inquiry modes may provide a framework through which past misunderstanding can be avoided. For example, qualitative inquiry also often involves a high priority on values inquiry. The focus in qualitative methods on "meaning" and on "individual experience" is at least in part a concern about the valuative component of sensemaking. In contrast, quantitative evaluators have traditionally eschewed systematic values inquiry in favor of the three inquiry modes that support representational sensemaking. At the least, the framework of inquiry modes may lead to more productive debate, in that

evaluators with different proclivities can explicitly discuss the merits of focusing different proportions of their efforts in support of the two general components of sensemaking. In addition, evaluators from a quantitative tradition may be more open to values inquiry when they see that it can be carried out systematically with methods familiar to them, such as surveys. On the other hand, qualitatively-oriented evaluators may have less resistance to quantitative approaches if it is clear that quantitative approaches can also meaningfully address values.

In addition to perhaps fostering greater understanding, the framework of inquiry modes also can facilitate communication by clarifying the differences between camps and by providing a common language for debate and discussion. With respect to the paradigm wars, qualitative methods, relative to quantitative methods, often (1) allow rapid, near-instantaneous iteration between inquiry modes and, (2) relatedly, allow decisions about which inquiry mode is primary to emerge at a later point. For example, Smith (1997) described an evaluation in which, in our language, she iterated fluidly between description, classification, and causal analysis. The ability to move between inquiry modes, and to delay decisions about which mode is most important in an inquiry, may be an important source of the appeal of qualitative methods.

The inquiry modes also can be used to identify important differences between alternative evaluation theories and approaches. For example, Wholey's approach to evaluation emphasizes description, in the form of monitoring. Campbell, in contrast, focused on causal analysis. To reasonably differentiate between evaluation theories, however, other aspects of our broader framework must also be considered. These other features are also important for evaluation planning. We turn now to other key features that need to be considered in conjunction with inquiry modes.

INQUIRY MODES IN CONTEXT

What should evaluators consider in choosing from among the four inquiry modes? Or when choosing from among the many options that exist for combining across inquiry modes? Or when they are choosing specific methods from within an inquiry mode? We cannot answer these questions in detail here, though it is possible to sketch out some of the key components of the answers. One important factor, to which we now turn, involves the broader purpose of an evaluation.

Social Betterment and Four Evaluation Purposes

Elsewhere (Mark, Henry, & Julnes, in press), we suggest that there are four primary purposes in service of which evaluation findings can be employed: (1) assessment of merit and worth, (2) oversight and compliance, (3) program and organizational improvement, and (4) knowledge development. This fourfold distinction builds on previous work by Scriven (1967), Chelimsky (1997), and Patton (1997). Although the four purposes are listed separately, an actual evaluation can serve more than one purpose. Nevertheless, we believe that most actual evaluations are conducted to accomplish one or two primary purposes.

The first evaluation purpose, *assessment of merit and worth*, refers to the development of a warranted judgment of the important characteristics and the value of a program or policy. Thus, assessment of merit and worth corresponds to what Scriven (1967) originally called

summative evaluation. We prefer the term "assessment of merit and worth" to summative evaluation, as it seems more descriptive. In particular, the term reminds us that judgments are to be drawn both of merit (i.e., the quality of the program or policy, in terms of whether it performs well) and of worth (i.e., the value that this brings to the larger social good).

The explicit reference to both merit and worth highlights some difficult and important issues. One involves the procedures for translating findings about merit (What does a program accomplish? What are its attributes, in particular in terms of rights and procedural guarantees?) into judgments about worth (How valuable is this for society?). Many previous writings about evaluation do not deal explicitly with this process (for an important exception, see Scriven, 1994). As noted above, we suggest that democratic processes and institutions are truly responsible for making judgments of worth based on findings about merit, but that evaluators can greatly aid in that process by reporting findings about merit in relation to findings from values inquiry.

The term *oversight and compliance* refers to assessments of the extent to which a program follows the directives of statutes, regulations, or other mandates. Oversight and compliance evaluations focus on such issues as whether the program services being delivered are the services that have been authorized (McLaughlin, 1975), whether program clients meet established eligibility criteria (GAO, 1998b), or what percentage of the target population is being served (GAO, 1998a). Such evaluations serve program sponsors', funders', and the public's needs for oversight and compliance: Is the program carrying out the activities it is supposed to do? Are funds being spent as authorized?[2] Although they can answer, in an operational sense, whether a program is following mandated practices, oversight and compliance evaluations do not in and of themselves give strong warrant for a determination of merit and worth. In many cases, a program may operate in full accord with legislation and regulations, and still not be effective. Nevertheless, it is sometimes important to know whether a program has been implemented as it is should be.

Program improvement, as a purpose of evaluation, refers to efforts to provide timely feedback designed to modify and enhance program operations. There are in fact somewhat different models of program improvement. In the approach suggested by Scriven, who coined the term "formative evaluation," program improvement is seen as involving a kind of mini-assessment of merit and worth. That is, a more casual and timely assessment of merit and worth is conducted, and the results are reported to program sponsors and staff who can then use the feedback to establish the need for program modifications. One probably underutilized technique for doing this is to rely on "natural experiments" that may suggest, for example, that one means of service delivery may be better than another.

Another approach to program improvement is analogous to the auto mechanic. The job of the mechanic is not to judge the merit and worth of a car relative to its alternatives, as *Consumer Reports* would do, but to make or suggest modifications that will likely enhance the car's performance. Evaluations of this form often focus on diagnosing the parts of the program that are causing observed problems, considering alternative approaches to these parts of the service delivery package, and deciding which change to pursue. Wholey's (1983) approach to evaluation, in which monitoring systems are developed for manager's use in revising program operations, illustrates this approach, as does much of the "quality improvement" movement.[3]

Yet another approach to program improvement, which overlaps the others conceptually, is model based. In this approach, one identifies some model to which, it is believed, the program should aspire. Often the model is described as involving "best practices." Any

discrepancy between actual program practices and the model (best practices) is seen as evidence of the need for program revisions. *If* there is actually good evidence that the best practice model has demonstrable merit and worth, then this is in fact a form of assessing merit and worth via classification. However, our sense is that very often the so-called best practices are actually in large part a subjectively determined set of things that specialists think *should* be related to desirable outcomes, even though good evidence is lacking. In such cases, we would label this as a form of program improvement effort, rather than as an assessment of merit and worth.

In addition, Chelimsky (1997) points out that some evaluators seem concerned, not so much with improving a specific program, as with improving organizational capacity to set policies, design and administer programs, and evaluate. We see this as a special type of program improvement, where the objective of program improvement is addressed from a broader, systems perspective.

Knowledge development refers to efforts to develop and test general theories and propositions about social processes and mechanisms as they occur in the context of social policies and programs. For some scholars, the world of social policies and programs is a valuable laboratory for developing and testing hypotheses and theories. The researcher interested in knowledge development may not be interested in the program or policy per se, but is using it primarily as an environment that allows the investigation of some research question of interest. Knowledge development can be a valuable adjunct to other evaluation purposes, and can in some cases make major contributions to social betterment.

Knowledge development can focus on a wide variety of research questions, and on large social science theories or "small theories" of local programs (Lipsey, 1993), depending on the researcher's interests. For example, scholars of social programs or of public administration might, by virtue of evaluation work done on social programs, attempt to develop general theoretical propositions about the implementation of social programs (Scheirer, 1987). Scholars interested in a particular area of human services, such as assisted living programs for the elderly, might try to develop a general classification system to describe the different types of services delivered in this area (e.g., Conrad & Beulow, 1990). Alternatively, one might attempt to develop a theory of the treatment types that are effective for different types of clients in a program area, as Lipsey (1997) has done in developing a general account of the relative effectiveness of different approaches to juvenile offenders. Or one might be involved with the evaluation of some policy or program because it allows the application of some newly developed research methodology.

Of course, in practice the four purposes are not mutually exclusive. If the evaluator is skillful enough, and resources suffice, multiple purposes can be addressed within a single evaluation or across a series of program evaluations. Nevertheless, making choices from among the four evaluation purposes should be among the first steps in evaluation planning.

Although we will not detail the linkages here, we believe that the selection of evaluation purposes should be based on an analysis, for that specific situation, of the extent to which each evaluation purpose can serve the relevant deliberations about the program or policy. In the context of social programs and policies, this means that evaluators must try to discern which of the evaluation purposes will best serve the democratic processes and institutions that are responsible for defining and striving for social betterment. Social betterment, that is, the alleviation of social problems and the meeting of human needs, is the ultimate objective of evaluation (for more on the concept of social betterment, see Henry & Julnes, 1998; Mark, Henry, & Julnes, in press). However, evaluation in the public sphere does not accomplish

social betterment directly. Instead, evaluation should provide assisted sensemaking to those democratic institutions and processes that wrestle with the questions of what betterment is and how to achieve it.

Purposes and Inquiry Modes as a Broader Framework for Communication

Earlier we noted that the framework of inquiry modes can perhaps facilitate communication, by allowing evaluators to see better the priorities and activities of those following some other approach to evaluation. The taxonomy of four evaluation purposes increases this potential contribution. By considering both inquiry modes and purposes, it is easier to see points of convergence (if any) and divergence between different schools of evaluation. For example, both Campbell and Scriven give priority to the assessment of merit and worth, but they emphasize different inquiry modes. Campbell (e.g., Campbell, 1969) focused on causal analysis, while Scriven (e.g., Scriven, 1990, 1994) appears to employ mostly a combination of description and classification (in terms of his checklist methodology, for example). Wholey (e.g, Wholey, 1983), in contrast, emphasizes program improvement, to be accomplished primarily through the descriptive inquiry mode, with the results from monitoring fed back to managers who are charged with enhancing program operations. In similar fashion, we believe it is both instructive and potentially conducive to improved communication to attempt to locate various evaluation approaches in terms of which evaluation purpose and inquiry mode they emphasize. Of course, no simple summary statement is likely to do complete justice to wide-ranging, thoughtful evaluation theorists. For example, Scriven (1976) has also identified the "modus operandi" method of causal analysis, whereby some recognizable signature implicates a particular cause. Nevertheless, we believe it is useful to identify the general tendencies of different evaluation approaches, in terms of evaluation purpose and inquiry mode.

A FRAMEWORK FOR PLANNING

The framework of evaluation purposes and inquiry families is, we argue, not just useful as a language for better communication across different camps in evaluation. It can also serve as the core of a framework for planning individual evaluations and series of evaluations. Although there are not any hard-and-fast one-to-one correspondences between evaluation purposes and the inquiry modes, some associations appear to have evolved in practice. For example, many of the approaches that emphasize the purpose of program improvement use the inquiry mode of description; Wholey's work, as noted previously, is a prime instance. That is, Wholey (e.g., 1983) generally emphasizes the use of monitoring systems (description) to provide information to managers to adjust program activity (program improvement). In contrast, when causal analysis is carried out in evaluation, most often it is in the service of assessing merit and worth.

A thorough discussion of the most common or the most preferred linkages between evaluation purpose and inquiry family is not possible here; instead we simply describe in this section, in summary form, some of the factors that should be considered in selecting among the evaluation purposes and inquiry families.

Selecting the Evaluation Purpose(s)

The primary consideration in the selection of an evaluation purpose (or purposes) should be the extent to which each purpose is likely to contribute to social betterment in that particular case. For example, are the democratic processes and institutions at a crossroads where decisions are about to be made about which of two programs to implement more widely, or about whether to continue a policy? If so, assessment of merit and worth would be called for. In other cases, a program may be on-going, stable, and unlikely to be replaced. In such an instance, the concern may be mostly with how services can be improved or costs lowered. Program improvement evaluations address these types of questions. One reason for the apparent rise in the frequency of evaluations that focus on program improvement may be that this purpose fits well with the needs that exist in very common situations, with programs that are stable and not likely to be replaced. Alternatively, there may be concerns about whether the program is being carried out as mandated, and about whether proper management procedures are in place. Such concerns would likely lead to oversight and compliance as the highest priority evaluation purpose. In these and other ways, the choice of evaluation purpose in the public sector should be based on judgments about the way each purpose can contribute to democratic processes and institutions.

Selecting the Inquiry Mode(s)

Inquiry mode(s) should be selected in light of evaluation purpose and related considerations. Purpose does not, however, drive inquiry mode in any simple one-to-one linkage. More than one inquiry mode can be used to address any single evaluation purpose. But this does not imply that "anything goes" when it comes to selecting an inquiry family. It is beyond the scope of this article to consider all of the possible linkages between evaluation purpose and inquiry mode. We can, however, illustrate the relevant reasoning with one purpose (the assessment of merit and worth) and two inquiry modes (causal analysis and description).

When the evaluation purpose is the assessment of merit and worth, causal inquiry is often used as an inquiry mode to obtain evidence about the effects of a policy or program. Does the program alleviate the social problem it was designed to address? More generally, what are its positive and negative effects, for whom, and under what conditions? (In many instances, causal analysis is not sufficient by itself, but we will not discuss these complexities here). Alternatively, description, in the form of monitoring, might be carried out, and may sometimes suffice as a rough assessment of merit and worth. In particular, outcome monitoring may demonstrate that outcome variables are not at a troublesome level, or that they are showing improvement.

As noted above, without the more targeted methods of causal analysis, monitoring will not allow a confident conclusion that the program, and not some other factor, is responsible for the level of the outcomes. However, in some circumstances, it can persuasively be argued that democratic processes and institutions may not require the more conclusive evidence of causal analysis. Sometimes it may actually be more useful simply to know whether, for the social problem that motivated the program, things are getting better. That is, monitoring may suffice to give a rough judgment of merit and worth, in the sense that society appears to be on the path to betterment, even without a relatively confident attribution of causation to the program. The worn expression, "If it ain't broke, don't fix it," conveys the sort of satisficing

logic (Simon, 1997) through which findings from monitoring may suffice as a rough assessment of merit and worth. For example, for those involved in a comprehensive community substance and violence prevention coalition, including the public, it may be sufficient to find the program meritorious if drunk driving, drug-related crimes, and violent crime are all down, and school students' and teachers' safety is up, even though the program may not be entirely responsible for these outcomes.

Choosing From Among the Choices

If more than one inquiry mode can serve a given purpose, how should the evaluator choose among them? Adding even more complexity, how are decisions to be made about whether to concentrate on one purpose, and do it very well, or to address more than one purpose, perhaps doing less well for each?

In short, no simple answers exist. Judgment is required. Tradeoffs abound. Moreover, truly sensitive answers to these and similar example are better if they can be given in a particular context. It is possible, nevertheless, to consider some general guidelines. These presume that there are defensible reasons for considering one option as better than the other. For example, as we have already suggested, for the assessment of merit and worth, causal analysis is usually better than the monitoring activities of description, because it is easier to attribute the outcomes to the treatment and not to some other factor. Alternatively, one may be forced with a judgment of attempting to facilitate program improvement with either an ongoing monitoring system, or with a one-time performance audit, and it may be judged that the ongoing system would be a better option. Under such circumstances, where there are options to be chosen from and the options can be ranked in desirability, one would probably consider several factors, including:

1. The relative cost of the two approaches—if the stronger option is as cheap as the weaker, there would seemingly be no reason to choose the weaker. In contrast, depending on resource constraints, one may be more likely to accept the weaker of two options if the stronger option is much more costly;
2. The cost of getting the wrong answer, and the likelihood that erroneous conclusions are greater with the weaker option—generally, one would be less likely to accept the weaker of two options to the extent that the costs of making the wrong conclusion are high;
3. The likelihood that, if the wrong decision is made, the consequences will be observed and can be reversed—one would probably be more willing to accept the weaker of two options to the extent that, if conditions subsequently declined, this decline would be observed and corrective steps could be taken; and
4. The quality of prior evidence—one would be more likely to accept the weaker of two options to the extent that there is already some strong prior evidence that seems to apply to the issue at hand.

For more discussion of these generic considerations, see Mark (1998). For more contextually-based—and therefore more useful—discussion of choices about evaluation purpose and inquiry mode, see Mark, Henry, and Julnes (in press).

Another important consideration, we believe, can help guide choices about inquiry

modes. This is the philosophy of science known as realism. Although there are a number of ways that realism can help guide the planning of evaluations, we note only one here.

The brand of realism that we endorse involves a spirit of openness toward both exploratory and confirmatory approaches, and particularly toward iterating between them. Realism suggests that iterating between exploration and confirmation is often needed to probe the complexities of an open system world with stratified levels. Elsewhere, we have discussed the needed iteration in terms of "competitive elaboration," "principled discovery," and their integration (Mark, Henry, & Julnes, 1998; Mark et al., in press). Space does not allow a detailed treatment of these concepts here. Suffice it to say simply that they involve an iteration between data and explanation, between induction and deduction, and between more exploratory and more confirmatory methods. While we believe that such iteration is almost always characteristic of outstanding evaluation work, its more widespread use can be encouraged, perhaps especially among those trained in a more quantitative tradition.

Realism also may play an important role in maintaining a lasting peace in the paradigm war, to which we referred at the start of this paper. Fittingly, then, we end this article with some brief comments about this potential role of realism.

REALISM: A POTENTIAL COMMON GROUND

Realism has a number of variants, with different theorists taking somewhat different positions about the nature of reality, truth, and human's ability to ascertain or approximate truth (Harré, 1986; Julnes, Mark, & Henry, 1998). Common to all realists, however, is the belief that there is an external reality out there. Experiences may be subjective. The way a person construes a situation matters, of course. But realists believe that the situation, and another person's construal, are "out there;" they are phenomena that have a real existence apart from one's own construction of them.

From one point of view, to advocate realism is quite unremarkable. We believe the term "evaluators" could be substituted into Meehl's comment about scientists:

> As to realism, I have never met any scientist who, when doing science, held to a phenomenalist or idealist view; and I cannot force myself to take a nonrealist view seriously even when I work at it. So I begin with the presupposition that the external world is really there, there is a difference between the world and my view of it, and the business of science is to get my view in harmony with the way the world really is to the extent that is possible (Meehl, 1986, p. 322).

In one sense, evaluators—like anyone else acting in the world—must at some level assume that the world is out there, spinning on its axis. Disbelief in an external world would seem to prohibit any semblance of normal social behavior.

From another point of view, realism offers a potential common ground for evaluators from both qualitative and quantitative traditions. For example, realism has a place for both qualitative and quantitative methods, and sees both as constructed technologies that have specific benefits and limitations. In addition, realism can offer a more sensible common ground, avoiding the excesses of the more extreme positions in the paradigm war (Julnes & Mark, 1998). Instead, realism offers an alternative. Realism supports the concern in evaluation for things such as the causal effects of programs, which can contribute to the representational component of sensemaking. Realism also can support concern with the

valuative component of sensemaking, including an interest in the subjective interpretations that clients and others have of these effects.

The American realist Hilary Putnam also offers some interesting insights that are relevant both to the role of evaluation as assisted sensemaking, and to the possibility of developing better communication within the field of evaluation. Putnam (1990) suggests that it is inappropriate to think of "solving" ethical problems in the same way that we think of solving mathematical puzzles or scientific problems: "The very words *solution* and *problem* may lead us astray. . . . I suggest that our thought might be better guided by a different metaphor—a metaphor from the law, instead of a metaphor from science—the metaphor of *adjudication*" (Putnam, 1990, p. 181). Putnam's view that ethical problems should be adjudicated is entirely consistent with our view that decisions about social policies and programs should be addressed through democratic institutions and processes.

Putnam also makes an interesting observation regarding the criteria necessary for adjudication in a democracy:

> To adjudicate ethical problems successfully, as opposed to 'solving' them, it is necessary that the members of a society have a sense of community. A compromise that . . . cannot pretend to derive from binding principles in an unmistakably constraining way, can only derive its force from a shared sense of what is and what is not reasonable, from people's loyalties to one another, and a commitment to 'muddling through' together (Putnam, 1990, p. 185).

A sense of community and a common language for discussion may be necessary for those who would consider and possibly act on our findings; so too may they be important for us who are evaluators. We hope that the framework summarized here might help us all to better "muddle through" together.

NOTES

1. An earlier version of this article was presented at the 1998 AEA annual meeting (Mark, 1998); this paper draws somewhat from Mark, M. M., Henry, G. T., and Julnes, G. (in press).
2. Despite their differences, oversight and compliance evaluations and those that assess merit and worth often arise from the same underlying motivation, that is, a desire to ensure that programs and policies are "delivering the goods." In the case of assessments of merit and worth, the "goods" in question are the valued outcomes that the program is supposed to bring about for its clients and possibly others. In the case of oversight and compliance evaluations, the focus is on ensuring that the program delivers the planned services to the intended clients. Perhaps because both purposes can arise from this same underlying motivation, past authors have tended not to differentiate between these two purposes (e.g., Chelimsky, 1997). However, the two types of evaluation do seem to represent distinct purposes and will typically require different inquiry modes, and so are differentiated here.
3. In this approach, the ongoing monitoring also then serves as an assessment of the merit and worth of the revision to the program, in a form of interrupted time series causal analysis.

REFERENCES

Bhaskar, R.A. (1975). *A realist theory of science*. London: Verso.
Bhaskar, R.A. (1998). *The possibility of naturalism* (3rd ed.) London: Routledge.

Bryk, A. S. (Ed.) (1983). *Stakeholder-based evaluation.* New Directions for Program Evaluation, no. 17. San Francisco, CA: Jossey-Bass.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24*, 409-429.

Caracelli, V. J. & Greene, J. C. (1997). Crafting mixed-method evaluation designs. In J. C. Greene & V. 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.

Chelimsky, E. (1997). The coming transformation in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: Sage.

Conrad, K. J. & Beulow, J. R. (1990). Developing and testing program classification and function theories. In L. Bickman (Ed.), *Advances in program theory*. New Directions for Program Evaluation, no. 74. San Francisco, CA: Jossey-Bass.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In. L. B. Sechrest & A. G. Scott (Eds.) *Understanding causes and generalizing about them*. New Directions for Program Evaluation, no. 57, San Francisco, CA: Jossey-Bass.

Cook, T. D. & Shadish, W. R. (1986). Program evaluation: The worldly science. *Annual Review of Psychology, 37*, 193-232.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

Datta, L. (1994). Paradigm wars: A basis for peaceful existence and beyond. In C. S. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives.* New Directions in Program Evaluation, no. 61. San Francisco: Jossey-Bass.

Datta, L. (1997). A pragmatic basis for mixed-method designs. In J. C. Greene & V. J. Caracelli, (Eds.), *Advances in mixed-method evaluation*. New Directions in Evaluation, no 74. San Francisco, CA: Jossey-Bass.

Flick, U. (1998). *An introduction to qualitative research*. Thousand Oaks, CA: Sage.

Harré, R. (1986). *Varieties of realism*. Oxford: Blackwell.

Henderson, L. A., Basile, K.C., & Henry, G.T. (1999). *Prekindergarten Longitudinal Study: 1997–98 School Year Annual Report*. Atlanta, GA: Applied Research Center, Georgia State University.

Henry, G. (1999). *What do we expect from preschool? A systematic inquiry into values, conflicts, and consensus*. Paper to be presented at the American Evaluation Association annual meeting, Orlando, FL.

Henry, G. T., Henderson, L. W., & Verrill, Linda. (1999) *Triggering mechanisms for school success: A theory based approach for evaluating preschool programs.* Paper presented at the American Educational Research Association Annual Meeting, Montreal Canada.

Henry, G. T. & Julnes, G. (1998). Values and realist evaluation. In G. T. Henry, G. Julnes, & M. M. Mark (Eds.), *Realist evaluation: An emerging theory in support of practice*. New Directions in Evaluation, no. 78. San Francisco, CA: Jossey-Bass.

House, E. R. (1991). Realism in research. *Educational Researcher, 20*, 2-9.

House, E. R. (1994). Integrating the quantitative and qualitative. In C. S. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives*. New Directions in Program Evaluation, no. 61. San Francisco: Jossey-Bass, 1994.

Julnes, G. & Mark, M. (1998). Evaluation as sensemaking: Knowledge construction in a realist world. In G. T. Henry, G. Julnes, & M. M. Mark (Eds.), *Realist evaluation: An emerging theory in support of practice.* New Directions for Evaluation, no. 78. San Francisco, CA: Jossey-Bass.

Julnes, G., Mark, M. M., & Henry, G. T. (1998). Promoting realism in evaluation: Realistic Evaluation and the broader context. *Evaluation, 4*, 483-503.

Lincoln, Y. S. (1990). The making of a constructivist: A remembrance of transformations past. In E. G. Guba (Ed.), *The paradigm dialog*. Newbury Park, CA: Sage.

Lincoln, Y. S. & Guba, E. G. (1994). RSVP: We are pleased to accept your invitation. *Evaluation Practice, 15*, 179-192.

Lipsey, M. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott

(Eds.), *Understanding causes and generalizing about them*. New Directions for Program Evaluation, vol. 57. San Francisco, CA: Jossey-Bass.

Lipsey, M. (1997). What can you build with thousands of bricks? Musings on the cumulation of knowledge in program evaluation. In D. J. Rog & D. Fournier (Eds.) *Progress and future directions in evaluations: Perspectives on theory, practice, and methods*. New Directions for Evaluation, vol. 76. San Francisco, CA: Jossey-Bass.

Marcon, R. (1992). Differential effects of three preschool models on inner-city four year olds. *Early Childhood Research Quarterly, 7*, 517-530.

Marcon, R. (1994). Doing the right thing for children: Linking research and policy reform in the District of Columbia public schools. *Young Children*, 2-10.

Mark, M. M. (1990). From program theory to test of program theory. In L. Bickman (Ed.), *Program theory in program evaluation*. New Directions for Program Evaluation, No. 47. San Francisco: Jossey-Bass.

Mark, M. M. (1998). *A conceptual framework for evaluation practice: A realist guide for planning evaluations in support of social betterment*. Paper presented at Evaluation '98: Annual Meeting of the American Evaluation Association. Chicago, November 1998.

Mark, M. M., Feller, I., & Button, S. (1997). Integrating qualitative methods in a predominantly quantitative evaluation: A case study and some reflections. In J. Greene & V. Caracelli, (Eds.), *Advances in mixed-method evaluation*. New Directions for Evaluation, No. 74. San Francisco, CA: Jossey-Bass.

Mark, M. M., Henry, G. T., & Julnes, G. (1998). A realist theory of evaluation practice. In G. T. Henry, G. Julnes, & M. M. Mark, (Eds.), *Realist evaluation: An emerging theory in support of practice*. New Directions for Evaluation, no. 78. San Francisco: Jossey Bass.

Mark, M., Henry, G. T., & Julnes, G. (in press). *Evaluation making sense of policies and programs: Description, classification, causal analysis and values inquiry*. San Francisco, CA: Jossey-Bass.

Mark, M. M., & Shotland, R. L. (1985). Stakeholder-based evaluation and value judgments. *Evaluation Review, 9*, 605-626.

McLaughlin, M. W. (1975). *Evaluation and reform: The case of ESEA Title I*. Cambridge, MA: Ballinger.

Meehl, P. (1986). What social scientists don't understand. In R. A. Schweder & D. W. Fiske (Eds.), *Metatheory in social science: Pluralisms and subjectivities*. Chicago: University of Chicago Press.

Melkers, J. & Willoughby, K. (1998). The state of the states: Performance-based budgeting in 47 out of 50. *Public Administration Review, 58*, 66-73.

Orwin, R. G., Sonnefeld, L. J., Cordray, D. S., Pion, G. M., & Perl, H. I. (1998). Constructing quantitative implementation scales from categorical services data: Examples from a multisite evaluation. *Evaluation Review, 22*, 245-288.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text*. Thousand Oaks, CA: Sage.

Pawson, R. & Tilley. N. (1997). *Realistic evaluation*. Thousand Oaks, CA: Sage.

Putnam, H. (1987). *The many faces of realism*. LaSalle, IL: Open Court.

Putnam, H. (1990). *Realism with a human face*. Cambridge, MA: Harvard University Press.

Putnam, H. (1994). *Words and life*. Cambridge, MA: Harvard University Press.

Reichardt, C. S. & Cook, T. D. (Eds.). (1979). *Qualitative and quantitative methods in evaluation research*. Beverly Hills, CA: Sage Publications.

Reichardt, C. S. & Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. In C. S. Reichardt & S. F. Rallis (Eds.), *The Qualitative-quantitative debate: New perspectives*, New Directions for Program Evaluation, no. 61. San Francisco, CA: Jossey-Bass.

Rosch, E. & Bloom, B. B. (Eds.). (1978). Cognition and categorization. Hillsdale, NJ: Erlbaum.

Scarr, S. Eisenberg, M., & Deater-Deckard, K. (1994). Measurement of quality in child care centers. *Early Childhood Research Quarterly, 9*, 131-151.

Scheirer, M. A. (1987). Program theory and implementation theory: Implications for evaluators. In L.

Bickman (Ed.), *Using program theory in evaluation*. New Directions for Program Evaluation, no. 33. San Francisco, CA: Jossey-Bass.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives on curriculum evaluation*. AERA Monograph Series on Curriculum Evaluation, no.1. Skokie, IL: Rand McNally.

Scriven, M. S. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual.* Beverly Hills, CA: Sage.

Scriven, M. S. (1993). *Hard won lessons in program evaluation.* New Directions for Evaluation: No. 58). San Francisco: Jossey Bass.

Scriven M. (1990) *The evaluation thesaurus*. Thousand Oaks, CA: Sage.

Scriven, M. S. (1994). The final synthesis. *Evaluation Practice, 15*, 367-382.

Sherif, M., Harvey, O. J., White, B. J., Bond, W. R., & Sherif, C. W. (1961). *Intergroup cooperation and competition: The robbers cave experience*. Norman, OK: University Book Exchange.

Simon, H. A. (1997). *Administrative behavior: A study of decision making processes in administrative organizations* (4th ed.). New York: Free Press.

Smith, M. L. (1994). Qualitative plus/versus quantitative: The last word. In C. S. Reichardt & S. F. Rallis (Eds.), *The Qualitative-quantitative debate: New perspectives*. New Directions for Program Evaluation, no. 61. San Francisco, CA: Jossey-Bass.

Smith, M. L. (1997). Mixing and matching: Methods and models. In J. C. Greene & 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, CA: Jossey-Bass.

Spradley, J. P. (1980). *Participant observation*. New York: Holt, Rinehart and Winston.

U. S. General Accounting Office (1998a). *District of Columbia: Extent to which schools receive available federal education grants*. GAO/HEHS-99-1. Washington, DC: Author.

U. S. General Accounting Office (1998b). *Head Start: Challenges in monitoring program quality and demonstrating results*. GAO/HEHS-98-186. Washington, DC: Author.

Weiss, C. H. (1983). The stakeholder approach to evaluation: Origins and promise. In Anthony S. Bryk (Ed.), *Stakeholder-based evaluation*. New Directions for Program Evaluation, no. 17. San Francisco, CA: Jossey-Bass.

Wholey, J. S. (1983). *Evaluation and effective public management.* Boston, MA: Little Brown.

Wholey, J. S. (1994). Assessing the feasibility and likely usefulness of evaluation. In J. S. Wholey, H. P. Katry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco: Jossey Bass.

Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds.). (1994). *Handbook of practical program evaluation*. San Francisco, CA: Jossey-Bass.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*, 749-750.

Yin, R. K. (1994) *Case study research* (2nd ed.). Thousand Oaks, CA: Sage.