The Virtual Reality Of Systemic Effects Of NSF Programming On Education: Its Profession, Practice, Research, And Institutions

Robert E. Stake
University of Illinois

It is both healthy curiosity and political necessity to wonder how research and development in science education is affecting not only the teaching and learning of science but also the greater educational and social system. In this paper, I review concerns about program effectiveness and accountability, and comment on the capabilities of program evaluation methods and people to trace systemic effects. Before identifying potential contributions from qualitative methodology, I outline its common characteristics. Claiming an interpretive commitment to be qualitative research's characteristic most applicable here, I suggest creation of, for each major program of the directorate, a semi-independent evaluation council for long-term interpretive study of the systemic influence of NSF educational research and development on various fields of action.

Seeking New Strategies for Program Evaluation

Thirty years of experience with the evaluation of Federal programs has persuaded many members of the American Evaluation Association that "there are no easy answers." At each year's annual meeting, there are restatements of the perplexity and renewed attention to political and cultural contexts. The foundation for future strategic thinking should not ignore the presidential addresses, the 96 theses of Lee Cronbach and colleagues (1980), and the 31 "hard-won lessons" identified by Michael Scriven (1993). Applying some of the experiential wisdom expressed in those resources to the present task, I begin with the following 17 caveats.

Evaluation Strategies: Caveats

1. Providing indicators of program impact is a task fraught with political and promotional pressure, resulting in overly "favorable" evaluations (Scriven 1991), resulting in evaluation schemes that exceed the technical capacities of evaluators. Realistic review of evaluation strategies is uncommon. Over-promising becomes routine. Organizational structures should be developed to require more realistic strategies for evaluating NSF programs.

2. Efforts to measure program merit and effect face complex political environments that reward:

   a. Delaying action (evaluation seldom can happen fast enough to support or counter impressions and experiences of the program itself);

   b. Disguise of advocacies (by reifying certain criteria of success and obscuring others, groups oriented to the reified criteria are supported); and

   c. A facade of accountability (the act of commissioning an evaluation makes it appear that the commissioning agency is acting responsibly).

New strategies need to be directed as much at disengaging evaluation from the advocacies of science and mathematics education as at finding new representations of effect.

"New strategies need to be directed as much at disengaging evaluation from the advocacies of science and mathematics education as at finding new representations of effect."

3. While group efforts to examine strategies for program evaluation should be encouraged, strategies are not necessarily strengthened by group consensus. Strength is also to be found in a diversity of ideas. It may be more important to add strategic options, some unpopular, to the armamentarium than to find a grand strategy that has few opponents.

4. Uniform strategies across programs is not an important end. Dissimilarity within and between programs requires nonuniform evaluation methods. If methods are too dissimilar, understanding of program effects will be low. With strategies too similar, unique contributions of individual programs will be understated (Cronbach, et al. 1980).

5. One strategy recognized almost universally is that multiple measures of important constructs are highly desirable. Conducting multiple studies is one way of getting multiple measures, some of which will help validate the constructs and others which will help illustrate the different interpretations given a construct in different settings.

6. Evaluation data can be newly generated by research or can be gathered from people who already are interpreting what is happening.

7. Most government-sponsored evaluations are designed in instrumentalist fashion; that is, the program is presented as an agent effecting some change in operations and productivity with certain benefit to a clientele. In the eyes of many advocates and clients, however, program quality is seen as the quality of services provided, as intrinsic quality rather than product quality. The social sciences are a reservoir of instrumentalist views; the humanities are a reservoir of intrinsic views. A review of evaluation strategies should consider both (Guba and Lincoln 1981).

8. Whether or not programs should be evaluated formally is a political and administrative matter more than a developmental and epistemological matter. It is common knowledge that formal evaluation studies have usually not provided critical input to government decision making about continuation or change in programs.

9. Evaluation occurs both formally and informally. Those closest to the scene tend to be more satisfied with informal than formal evaluation. People at a distance, especially those dubious about the program, tend to prefer formal and independent evaluation.

10. Most programs supported by the National Science Foundation are complex. Instruction and other discourse affected by NSF programs are simultaneously being affected by many other influences. Attribution of effect to NSF programs is problematic, at best.

11. The more distant an intended effect is from program activity, the more difficult the attribution. Distance can be a matter of time, place, personal interaction, content, or conceptualization.

12. The pre-announced metaphor of "footprints" as an indicator of effects of a program's passing should be given no more than a moment's thought. That metaphor raises the image of pristine sur-

faces, such as newly waxed floor or fresh sand at the beach, and the fitting of a slipper to Cinderella-like program agents. Real surfaces are scuffed, trammelled, and exposed to countless footfalls, and real programs rarely leave distinguishing marks. But the major flaw in the metaphor is its romantic image of an indicator that requires little human interpretation.

13. Education and human beings are extremely complex. We seldom can measure effects of educational research and development directly. Validity of measurement tends to diminish, the more indirect the indicator. For a nation, a school, and sometimes even a child, our indicators of program effect are quite indirect. Many are of low validity. Indicating the systemic effects of NSF programming on research, training, professional communication, and popular discourse directs attention to quite indirect outcomes.

14. We have indicators of high validity and those of low (Shavelson, et al. 1987; Guiton and Burstein, 1993). Misleading evaluations result from interpreting indicators beyond the limits of their validity. For example:

15. Indicators have a reactive effect. To get better test scores or other marks, schooling is redirected to better affect the indicator variable, sometimes at the expense of the real targets of education. Both insiders and outsiders increasingly substitute the indicator variable as the definition of education. Were we to create valid indicators of systemic effects, advocates and adversaries would probably find ways to subveil them.

16. Essentially, in evaluation studies, we are not aiming as much to identify program effects as to identify the value of the effects (Scriven, 1991). Value of effect requires consideration of costs. (In education, worth and costs are seldom measured in dollars.) At least as hard to measure as effects, values and cost measurements are seldom included in an evaluation design. Strong measurement designs often presume that values and costs will be apparent without measuring. Sometimes the best strategy will be to obtain summary judgments from people who themselves have been exposed to all three.

These indicators:	**are a good indicator of:**	**but a poor indicator of:**
need statements	what people would like	what is actually needed
standardized test scores	student ability	actual student achievement
grade point averages	compliance in instruction	ability to use own knowledge
courses taken in Education	teacher formal qualification	teaching quality
monetary costs	money spent	the social costs
followup ratings	participant satisfaction	program effectiveness

17. Increased attention is being given to the design of indicators of provision of educational opportunity. School delivery standards (Porter 1993) would change evaluation strategy to concentrate more on the measuring of process and less on the measuring of product. A strategy emphasizing systemic effects runs counter to emphasis on provision of opportunity.

"For most people, the evaluation of Federal programs raises the expectation that something will be measured to which a value can be attached."

I open my paper with these 17 caveats intending to help anchor discussions of evaluative strategy in practical experience. I think it is possible to increase NSF sensitivity to the effects its programs are having, but precise, validated, and immediate indicators are some of the illusory "easy answers." How NSF sensitivity and program advocacy may be enhanced by nontraditional evaluation strategies requires a careful look at what is expected of program evaluation.

What Is Being Asked of Evaluation

Essentially, evaluation is the determination of merit and shortcoming (Scriven, 1967). Program evaluation is commonly taken to be "systematic examination of events occurring in and consequent on a contemporary program ... to assist in improving this program and other programs having the same general purpose" (Cronbach, et al., 1980). For most people, the evaluation of Federal programs raises the expectation that something will be measured to which a value can be attached. (In this paper, I am not speaking of project or proposal evaluation but the evaluation of large NSF programs, especially their effects on the educational R & D enterprise and on education generally.)

A Contrived Rationality- Program evaluation, like the social sciences, is in the business of making rational what is empirical. Our principal knowledge of life is empirical. Although indirect and sporadic, much of our knowledge of the work of government is empirical. We try to rationalize what we experience. Government programs change, society changes, people change, all calling for changes in our rationalizations.

Evaluation specialists get contracts to discern a program's measurable relationships, particularly cause and effect relationships. And most evaluators confidently try—operating under the notion that if change has occurred, a cause can be discerned. If subsequent conditions seem to connect back to the program more than to anything else, then it may be said that the program caused the effects. Proof of such a relationship is far beyond reach. Certainly, in program evaluation, if not everywhere, cause and effect is a constructed reality—sometimes a contrived reality.

The context of government programs is political. Information needs are unlike contexts common to researchers (Chelimsky, 1991). Problems are real and taken seriously, but expediency and irrationality are common. Almost every government official is tuned to the morning news (Barnouw, 1970). Bureaucracies strive for rationality; failing that, for the appearance of rationality. They are beset by news media not only for news but for stories. The media are presumptuous about rationality. They equate rationality with responsibility. They imply rationality to be the responsibility of officials, whose information systems are expected to tell what is causing what.

Reporter orientation to causality is particularly aroused by a calamity such as the immolation of the Branch Davidian cult in Waco, Texas. Did the

FBI provoke a mass suicide? Did the President really take full responsibility? Looking back on the Waco calamity, columnist Michael Kelly of the Washington Post discerned the discrepancy between public and media stances, noting little interest within the public in finding someone to blame (April 1993). Kelly used words of Robert Coles, which described the media's "... arrogant faith in rationalism ... , all of them paying homage to the great delusion of our times, that social scientists will deliver us from irrational madness and the random hand of fate." Blame makes a good story. Under media expectations, it behooves evaluators to identify blame for program shortcomings.

Deliverance also makes a good story. Within professional education at present, much attention is paid the Curriculum and Evaluation Standards for School Mathematics, published by the National Council of Teachers of Mathematics (NCTM) in 1989. Does problem-solving get graduates ready for the work place? Is NCTM now leading the school reform movement? Some believe evaluators should be trying to measure such effects. How should they evaluate the effects of NCTM Standards? Perhaps by looking into other teaching areas (Ball, 1992). Specialists in language arts promoting a "whole language" approach occasionally mention the NCTM Standards. Specialists in distance education trying to develop simulations far from campus occasionally mention the NCTM Standards. Is their work influenced by the Standards? Possibly, but not on the basis of how frequent is the mention or how congenial the innovation. Workers in other fields see that the legitimacy of the Standards might rub off on their efforts, so they cite them. Citation does not mean they have been influenced by the mathematics teachers.

Now that we have thought about it, there may be a phenomenon we can call the NCTM effect on school improvement. And an evaluator might be able to estimate how much the work of mathematics teachers has influenced other innovatory efforts. Could we call the estimate an indicator? Could we validate the estimate? Indicator validation is not going to happen. The estimate itself may be useful, not only for promotional purposes, but in the rumination and discourse of program management. But estimates are not facts. Indices such as "the NCTM effect" or "readiness for the work place," just like the now vernacular "employment rate" and "Dow Jones average," however useful, are fictions, beyond constructed realities, a form of that new whiz bang, "virtual reality." More on that in a moment.

The real work of educators is not "to look good," nor is it "to catch up with the Japanese," nor is it even (in my view) "to cause the child to be different," but to provide opportunity and pressure for children to follow preferred paths to becoming educated. It is the natural state of the child to be affected by teachers and tenuously by distant research programs. How much the separate layers of the system can take credit for good effects—or bad, for that matter—is beyond the understanding of everyone, including evaluation specialists. Whatever the appetite for indicators, whatever the demand for program accountability, however useful measurement of effect might be, the state of the art is such that indicators of systemic effect are not available. And it is irresponsible for officials to use unvalidated indicators of effect as if they had been validated. And it is an act of deception for evaluators to provide such indicators.

"There is no single wellspring of qualitative research from which to draw methods for evaluating NSF strategies."

What state-of-the-art evaluators can do is to see if programs are drawing upon the best of human understandings, organizing programs in felicitous ways, recognizing and coping with problems, maintaining a dignity of human relations. It is not wrong to be curious about outcomes, but it is wrong to join in the deceit that governments cause education, and in the self-deceit that evaluators reliably measure and attribute effects. It is wrong to portray a rationality that does not exist.

It is also wrong to base evaluation strategy on what ought to be rather than on what is. Formal evaluation expectations are based largely on specialist services presently available. They do evolve, and can be seen to be increasing their use of qualitative field work, particularly with case studies and ethnographic interpretations. How NSF sensitivity and program advocacy may be enhanced by nonresidential evaluation strategies requires more than a passing knowledge of qualitative research methods. Drawing upon the *Handbook of Qualitative Research*, (Denzin and Lincoln, 1994), the following section is my distillation of that emerging methodology—disciplined qualitative inquiry.

The Nature of Qualitative Research

There is no single wellspring of qualitative research from which to draw methods for evaluating NSF strategies. Practices vary in different fields. The distinction between quantitative and qualitative methods is a matter of emphasis more than a matter of boundary. In each ethnographic or naturalistic or phenomenological or hermeneutic or holistic study, i.e., in each qualitative study, enumeration and recognition of differences in amount have a place. And in each statistical survey and controlled experiment, in each quantitative study, natural-language description and researcher interpretation are expected. Perhaps the most important differences in emphasis are threefold:

a. Distinction between knowledge discovered and knowledge constructed;

b. Distinction between aiming for explanation and aiming for understanding; and

c. Distinction between personal and impersonal roles of the researcher.

Constructed Knowledge and Virtual Representations- The children of today are manyfold the linguists their parents were as children. Their exposure to images has grown a hundredfold. Their access to keyboards and software gives them vast new ranges of expression. Literary empowerment has been enormous for evaluators as well. We can say so much more, represent it in so many more ways, prepare handsome camera-ready copy ourselves.

As the electronic field has exploded in both sophistication and public access, the art of representation has exploded too. Readers can be immersed in the description, drawn into the most elaborate of vicarious experiences (Spiro, et al. 1987). Following Aldous Huxley's *Brave New World*, broadcast advertising (Barnouw, 1970), and, more recently, computer artist Myron Krueger's Artificial Realities (1983), we are passing into a period of interactive stimulation that extends personal experience far beyond the movies and charismatic teaching. Among its champions, it is called, "virtual reality" (Woolley, 1992), making possible a sensory reality beyond

ordinary experience, such as playing tennis on a low gravity court. Radio talk shows have been titillating the public with ideas about simulating pleasure. A few "virtual reality" venues are more sober, more intellectual. A number of our colleagues in artificial intelligence research have designed extra-responsive environments for simulation of problem situations to enhance learning (Psotka, 1993). But this medium is one of grand deception. As Lewis Carroll explained, "For the snark was a boojum, you see."

What I said two paragraphs back about empowerment of children and evaluators is merely an assertion, another virtual reality, but one I expect will sit comfortably with most readers. If that claim is not true, it is virtually true. It is an untruth most people will accept as true. Increasingly we realize our dependence on virtual truths. We pause in our own metamorphosis. As we increase our ability to represent the world, we have greater difficulty in remembering what the world actually has been, and increasing doubt we ever knew what it might be. Some virtual realities we settle for, some we aspire to, such as those we call science and art. We cannot even imagine a world without these virtual realities, these constructs, these indicators. Our problem is one of believing them too much, losing the appetite for validation.

Multimedia shows and role playing repeatedly have shown us that simulation creates a reality of its own. When simulation is effective, that which was simulated can become secondary to the simulation. Shakespeare and McLuhan agreed, "The show's the thing." Virtual sunsets outdo the real in so many ways. The classical questions reappear: "What is reality?" "Is there substance behind appearance?" Children and researchers create new knowledge. And in telling others what they have learned, even as they remember, they simulate that knowledge. New knowledge and simulated knowledge are different (Stake and Trumbull, 1982), propositional and tacit knowledge are different (Polanyi, 1969), but I often find them difficult to tell apart.

In our personal lives, some symbols, narratives, and indices stand for the real thing, more stand for other symbols, narratives, and indices. We remember, sometimes remembering memories rather than the original experience. We create within our minds a world of representations. We do this from our earliest ages, seeking to make sense of puzzling environments, repeating experiences, refining our indicators—but all too seldom do we go out of our way to validate them.

In our societal and institutional lives, we of course need symbols, narratives, and indices. We do not know how to survive without them. We are jolted by the realization that indices are created for other purposes than representation: as dreams and icons, as subterfuge, as enhancements and caricatures, as provocations and supplications. Secretary of Education Terrell Bell created his famous Wallchart of SAT scores ostensibly to represent the quality of secondary education in the 50 states. He knew the data were greatly misleading, but posted them to provoke researchers into creating a valid comparison (Bell, personal communication). Indices exist for advocacy as much as for information. New indices are seldom validated over a developmental period before being offered for public or specialist interpretation. It is part of our evolving language, part of our evolving knowledge base, to have grand indices, but it is part of our carelessness to take them to mean what they seem to mean.

"As we increase our ability to represent the world, we have greater difficulty in remembering what the world actually has been, and increasing doubt we ever knew what it might be."

A preponderance of qualitative methodologists are constructivists, professing belief that knowledge is the invention of inquiring minds, not their discovery (Schwandt, 1994). Knowledge is made, not found. Qualitative study of teaching and learning correspondingly emphasizes the construction of ideas by children rather than the acquisition of ideas. This is not just a preferred set of learnings or preferred pedagogy, but an epistemological definition. Each person constructs knowledge, most not recognizably unique, but individually created. We have common knowledge not so much because there are pre-existing facts, truths, for us to discover, but because learning, like dressing and driving, is a social process. We have strong tendencies to conform. We modify our actions to fit the actions of those we respect. And we create knowledge that appears to be very similar to that of the people around us.

The important thing to the qualitative researcher is that it is helpful to consider much learning, much "reality," as human construction. It is necessary sometimes to be reminded that the indices, the virtuals, we use to monitor our lives are contrivances regularly in need of calibration.

Experiential Understanding- A distinction among aims, an epistemological distinction, fundamentally separates qualitative and quantitative inquiry. The distinction is not that between quantitative and qualitative data. The distinction is in intent, between inquiry for making explanations versus inquiry for promoting understanding. It has been nicely stated by philosopher George Henrik von Wright in his book, *Explanation and Understanding* (1971). Von Wright recognized that understanding is personally constructed. He acknowledged that explanations are intended to promote understanding and understanding is often expressed in terms of explanation—but epistemologically, the two are quite different. Von Wright emphasized the difference between generative explanation and experiential understanding.

It is a distinction seen in preferences for process versus product evaluation. Products are the manifestation of generative processes. Choosing product evaluation is problematic for us because the causes of systemic effects are not necessarily the processes we assume, allude to, or experience. Given such uncertainties, the qualitative evaluator gives greater attention to process as experienced (Guba and Lincoln, 1982), with the reader expected to share in the interpretation. For the educator, the distinction parallels the difference between preparing to teach didactically and preparing experiential opportunities for learners. Shall researchers tell a reader what they have learned, or shall they arrange a situation optimally suited to reader learning? Qualitative evaluation designs generally aim to have evaluators make descriptions and situational interpretations of phenomena, which they offer colleagues, students, and others for modifying their own understandings of program merit (that is, for "naturalistic generalization," as Deborah Trumbull and I called it in 1982). Quantitative evaluation designs generally aim to advance abstract comprehensions of the evaluators who, in turn, present these explanations to their colleagues, students, and diverse audiences.

Qualitative descriptions are expected to be recognizable by readers, yet no description captures veridically the phenomena described. Jorge Luis Borges spoke of this elusive character of language in *A Yellow Rose:*

> ...Then the revelation occurred: Marino saw the rose as Adam might have seen it in Paradise, and he thought that the rose was to be found in its own eternity and not in his words; and that we may mention or allude to a thing, but not express it...

Borges recognized the inescapable artificiality of description.

Quantitative research methods have grown out of search for grand theory. To establish generalizations that hold over diverse situations, most social science-oriented researchers make observations in diverse situations. They try to eliminate the merely situational, letting contextual effects "balance out." They try to nullify context in order to find salient and pervasive explanatory relationships. Qualitative evaluators treat the uniqueness of individual contexts as important to understanding.

Most program evaluation work has been dominated by science's search for grand explanation. Employment of formal measurement and statistical analysis, i.e., quantification, has occurred in order to permit aggregation of a large number of dissimilar cases, thus to position the researcher to make formal generalizations about the program. The appropriateness of scientific explanation for program evaluation has been questioned by Scriven (1978) and Cronbach (1980, et al.) on the grounds of the particularity of the evaluand, its situationality, and its political context. Both of them have emphasized the evaluator's responsibility for authoring program-specific descriptions and interpretations. Practicing evaluators draw upon both quantitative and qualitative methods, choosing one or the other to provide scientific explanation or experiential understanding.

Emphasis on Interpretation- Qualitative evaluation specialists such as Elliot Eisner (1979) and Egon Guba and Yvonna Lincoln (1981) have urged reliance on direct interpretation of events more than on interpreted measurement of attributes. All research designs feature interpretation—but, with standard quantitative designs, there is effort to constrain interpretation during that period extending from design of the study to analysis of the data. Standard qualitative designs call for the persons most responsible for interpretations to be in the field during that period, responding to the situation (Stake, 1975), making observations and interpretations simultaneously.

The difference is epitomized by two kinds of research questions. In quantitative studies, the research question typically embodies a relationship among a small number of variables, e.g., "Is there an enduring correlation between applicability of technological support and teacher qualification over a variety of situations?" Efforts are made to operationally bound the inquiry, to define the variables, and to minimize the importance of interpretation until data are analyzed. At the very beginning, it is important to anticipate how relationships between variables would reduce weakness in explanation and, at closing, it is important for the researchers to modify their generalizations about the variables. In between times, it is important not to let interpretation change the course of the evaluation study (Stake, 1994).

In qualitative studies, the research question typically orients to cases or phenomena, seeking patterns of unanticipated as well as expected relationship. For example, "What will happen to collegial relationships among teachers

"Practicing evaluators draw upon both quantitative and qualitative methods, choosing one or the other to provide scientific explanation or experiential understanding."

working with this program if all are obligated to emphasize a problem-solving pedagogy?" Or if the project had been implemented sometime in the past, "What happened?" Dependent variables are seldom operationally defined, situational conditions may not be known in advance, even the independent variables are expected to develop in unexpected ways. It is important to have the interpretive powers of the research team in immediate touch with developing events and ongoing revelations, partly to redirect observations and to pursue emerging issues. The allocation of resources is different. Reliance on carefully developed instruments and redundancy of observations typical in a quantitative study give way to placement of the most skilled researchers directly in contact with the phenomena and making much more subjective claims as to the meanings of data.

"Thick description, alternative interpretations, 'multiple realities,' and 'naturalistic generalization' are not only common; often they are aims for these nontraditional research methods."

In his fine summary of the nature of qualitative study, Frederick Erickson (1986) claimed that the primary characteristic of qualitative research is interpretation. He said that findings are not just "findings" but "assertions." Qualitative study is not alone in personalizing interpretation. Speaking of all social science, Henry Aaron (1978, 156) claimed:

> Outsiders may be lulled into thinking that issues are being debated with scholarly impartiality, when in fact more basic passions are parading before the reader clad in the jargon of academic debate.

Qualitative methods invite personal reflection. With intense interaction of researcher and actors in the field, with a constructivist orientation to knowledge, with sensitivity to participant intentionality and sense of self, however descriptive the report, the qualitative researcher expects to express personal views.

Erickson drew attention to the ethnographers' traditional emphasis on emic issues, those concerns and values recognized in the behavior and language of the people being studied. Geertz (1973) called it: "thick description." And often the aim is not veridical representation so much as stimulation of further reflection, optimizing readers' opportunity to learn. Stake and Trumbull (1982) called it "naturalistic generalization," a concern for assisting the reader's further understandings. It draws from history, philosophy, and literature, sometimes paralleling the artist's work. Claude Debussy, on composing La Mer, not at sea, but in his Paris studio, said:

> I have my memories and they are better than the seascapes themselves whose beauty often deadens thought. My listeners have their own store of memories for me to dredge up.

The function of research is not always to map the world but to sophisticate the beholding of it.

Thick description, alternative interpretations, "multiple realities," and "naturalistic generalization" are not only common; often they are aims for these nontraditional research methods. Such pursuit of complex meaning cannot be just designed in or caught retrospectively (Denzin and Lincoln, 1994). It seems to require continuous attention, an attention seldom sustained when the dominant instruments of data gathering are objectively interpretable checklists or survey items. An ongoing interpretive role of the researcher is prominent in the work of qualitative research.

Other Characteristics of Qualitative Research- In addition to its orientation away from cause-and-effect explanation and toward personal interpretation, qualitative inquiry is distinguished by its

emphasis on holistic treatment of phenomena (Schwandt, 1994). I have remarked already on the epistemology of qualitative researchers as existential (as opposed to causal or generative) and constructivist. These two views are correlated with an expectation that phenomena are intricately related to many coincidental actions and that understanding them requires a wide sweep of contexts: temporal and spatial, historical, political, economic, cultural, social, personal.

Thus the case, the activity, the event, is seen as critically unique as well as common. Understanding it requires an understanding of other cases, activities, and events. Uniqueness is recognized not primarily by comparing cases on a number of variables—there may be few ways in which this immediate case strays from the norm—but the collection of features, the sequence of happenings, is seen by people close at hand to be in many ways unprecedented and important; that is, a critical uniqueness. Readers are drawn easily to a sense of uniqueness as they read narratives, vignettes, experiential accounts (van Maanan, 1988). The uniquenesses are expected to be critical to the understanding of the particular case.

For all their intrusion into habitats and personal affairs, qualitative researchers are non-interventionists. In the field, they try not to draw attention to themselves or their work. Other than positioning themselves, they avoid creating situations to test their hypotheses. They try to observe the ordinary and they try to observe it long enough to comprehend what, for this case, ordinary means. For them, naturalistic observation has been the primary medium of acquaintance. When they cannot see for themselves, they ask others who have seen. When formal records have been kept, they scrutinize the documents. But they favor a personal capture of the experience, so they can interpret it, recognize its contexts, puzzle the many meanings, while still there, and pass along an experiential, naturalistic account to allow readers to participate in some of the same reflection.

Recognition of Risks- Qualitative study has everything wrong with it that its detractors claim. It is subjective. The contributions toward an improved and disciplined science are slow and tendentious. New questions are more frequent than answers. The results pay off too little in the advancement of social practice. The ethical risks are substantial. And the costs are high.

The effort to promote a subjective research paradigm is deliberate. Subjectivity is not seen as a failing to be eliminated but as an essential element of understanding. Still, personal understanding frequently is misunderstanding, by actors, by the researchers, and by readers. The misunderstanding may occur because of the intellectual shortcomings of the interpreter or because of weakness in protocol which fails to purge misinterpretation. Qualitative researchers have a respectable concern for validation of observations, they have routines for “triangulation” (Denzin, 1989) that can approximate in purpose those in the quantitative fields, but they do not have the protocols that put subjective misunderstandings to a stiff enough test.

Many phenomena studied take long to happen and evolve along the way. Often we need a long time to come to understand what is going on. The work is labor-intensive and the costs are hard to trim. Many of the studies are labors of love. Many findings are esoteric. The

worlds of commerce and social service benefit all too little from investments in these formal studies. The return may be greater for those who study their own shops and systems by these methods, but self-study so seldom brings the disciplined interpretations of the specialist into play.

Many qualitative studies are personalistic studies. The cares of observed human beings insinuate into issues of the present research. Privacy is always at risk. Entrapment is regularly on the horizon, as the researcher, although a dedicated noninterventionist, raises questions and options not previously considered by the respondent. A tolerable frailty of conduct nearby becomes a questionable ethic in distant narrative. Some of us "go native," accommodating to the viewpoint and valuation of the people at the site, then reacting less in their favor when back again with academic colleagues (Stake, 1986).

"It is not that we need more than a single indicator; it is the idea of indicator that is insufficient."

It is not simply a matter of deciding whether the gains in perspective are worth these costs. The attraction of intensive and interpretive study are apparent, and were earlier when qualitative designs were considered unworthy of respect by many research agencies and faculties—as by some, they still are. Researchers inquire. They are controlled by the rules of funding and their disciplines, but those influence how they will report their use of qualitative methods. All researchers use them. There are times when each researcher is interpretive, holistic, naturalistic, and uninterested in cause. Then, by definition, she or he is a qualitative inquirer. Administrators, too, have these leanings and use these methods. The question here is how disciplined concentration of these methods might improve the evaluation of systemic effects.

A Qualitative Strategy

Human Surveillance of Policy- One implication of qualitative methodology is to raise a caution flag on the use of "indicator variables"; yes, on all formal representations of complex phenomena. More than an intensive search for the closest indicator of an expected effect, we need disciplined scrutiny of this particular notion of effect. Interested in the effects of a research program on public policy, we may seek already-existing traces and we may create new indicators of changes in policy, but we should also extensively and repeatedly examine our conceptions of the research program and the public policy. Experimentalists (Boring, 1950) used to call it, "the criterion problem," the suitability of the representation.

As we first identify a program and a criterion policy, almost immediately we have expanding conceptualizations of the problem, the remedy, the effects. We have no single construct to represent, no true substance to indicate. It is not that we need more than a single indicator; it is the idea of indicator that is insufficient. We evaluators need to realize that we are asked for, and we ourselves yearn for, artifice, the hypothetical, the illusory. We propose indicators of things that do not exist other than in our imaginations. Many of the things we would indicate—the well-being of a child, the coherence of a curriculum, the fiscal integrity of a school district, the merit of a research policy—do not exist other than as mental contrivances. They are not things we can approximate. There is no way that we can test the validity of such "representations."

That does not mean we should purge our thoughts of indicators. We have no choice. Words are indicators, photographs are indicators, memories are indicators. We cannot communicate without representations of both the tangi-

ble and the intangible. Of course we will have indicators, not only in common discourse, but in all means of technical representation. The big question is how we will treat our indicators. Particularly, will we set them up as approximates to imagined truths, as substitutes for human sensitivity, for decision making? Will we use them to regulate our affairs?

Sometimes we will. We use various servomechanical systems: thermostats, cost-of-living increases, sliding scale cutting scores for admission. All, we hope, are subject to petition and override, but they are a part of our human systems. Some serve us well. Sometimes we wonder if they serve us well enough. The more the decisions impact indirectly on personal well-being, on differences in privilege, on the common good, the more we should worry that the indicators may be unwell and the more we should insist upon calibration in the form of close human surveillance.

It sometimes is supposed that a qualitative approach is fundamentally an aggregation and quantified analysis of data gathered in an qualitatively interpretive fashion (Miles and Huberman, 1984; Yin, 1989). While that may be useful, an essentially qualitative strategy for monitoring the effects of research is typified not by the establishment of quantitative indicators of qualitative phenomena, but by the establishment of disciplined and reflective human surveillance over all indicators, qualitative and quantitative.

These humans, these discerning humans, will use existing indicators and construct new ones. They will use multiple indicators to reflect the complexity of the phenomena and different perspectives found among people affected. They will couch their thinking and presenting of indicators in the language of experience, frames of time, place, and personality. If they do their work well, they will be a deterrent to overinterpretation of indicators, to the oversimplification of problems and solutions. They will demystify.

But they also will mystify. They will try to convey the best of insights of those who have most closely studied the matter. They will introduce new constructs, new models, and new scales. If they do their work well, they will not make it easier to command understanding, nor to make decisions. What they will offer is not indicators but virtuals, representations not of something real but essences of things understood. They will continue to remind us of the construction of our knowledge.

Interpretation Roles- Of the three pervasive characteristics of qualitative research I elaborated earlier, the most promising for extending NSF program evaluation is, I believe, interpretation. Interpretation is not a stranger at NSF, but more comprehensive and protected roles can be imagined. To come to understand the effects of major NSF programs, the qualitative strategy I propose is simple: an invigoration of interpretive responsibility, a mobilization of interpretive assets, an elevation of interpretive capability. I am echoing the plea of Cronbach and associates who called for much more vigorous collegial review of evaluation research (1980). The National Science Foundation needs comprehensive interpretation of what its science education programs are accomplishing (Katzenmeyer, 1993). The best contribution of qualitative methodology to such evaluation would be, I think, to enhance the role of systemic interpretation.

Individual evaluation studies aggregate poorly (Cronbach, et al. 1980), in

"My suggestion here is ... for one group, an institutional council, to review science education performances of importance to NSF, including the systemic effects of its programs."

NSF as elsewhere. Policy makers do not get the support they need. Program officers and individual evaluation contractors provide too little in the way of historical perspective and independence. To get independent views of quality, evaluators are sought who have little to gain or lose by the conclusions they draw. These people usually have but cursory knowledge of present and past operations. To enrich formal evaluation with existing knowledge of present and past operations, an evaluation assignment often goes to prior funded parties (and potentially future award winners) or their associates, but these people are pressured by personal and institutional relationships to constrain their inquiries. There are natural constituencies of researchers for curricular issues, technical advances, teacher training, and special pedagogies, each capable of providing traditional reviews of research, development, and evaluation studies, but more narrowly defined than the panoramic responsibility for science and mathematics education. Most advisory panel members lack the purview, independence, and time to provide historical perspective.

An Interpretation Council- One possible move would be to create within each NSF program or in the agency as a whole, an Interpretation Council, a small, full-time, internal but independent, evaluation-oriented policy-analysis team. Among the members should be persons well experienced in program evaluation, research integration (Cook, et al., 1992) and qualitative field study (Strauss and Corbin, 1990). Maintaining knowledgeable but dispassionate status would not be easy. Interpretation roles and council status would take time to develop. Although the appointments might be as difficult as those to the Supreme Court, the needed talents already exist among those who staff the Education Directorate. Members should be committed to gaining a thorough and enduring acquaintance with key issues, major projects, and related programs, yet having little vested interest in particular ones. This council should not replace the External Expert Panel, a more removed group needed for their interpretations (Katzenmeyer, 1993).

On the organization chart, the council perhaps should be a permanent free-standing affiliate, possibly attached to the Inspector General's office. Although much smaller, in some ways it would mimic the Government Accounting Office. GAO serves the Congress; the Council would serve an NSF program—but to provide interpretation and review rather than to complete studies. Like GAO, the Council should be obligated to stay relevant to the sweep of institutional responsibility, subject to multiyear mission renewal, and free to design and conduct individual program reviews. Even though dedicated to its sponsor, the Congress, GAO appears to me to have sufficient independence for designing studies, for occasional unwelcome findings, and for initiating some inquiries unrequested (Chelimsky, 1987). With strong management and a capable staff, I would say that presently GAO is the outstanding program evaluation shop in the world today. GAO is not an ideal model, however, because it does not maintain a sufficiently enduring relationship with particular programs. The purpose of that agency is not long-term administrative reflection and continuing program evaluation.

Thomas Cook (1978) and Lee Cronbach and associates (1980) pointed to the desirability of "social problem study groups." My suggestion here is similar but different. It is for one group, an institutional council, to review sci-

ence education performances of importance to NSF, including the systemic effects of its programs. One organizational model to examine would be the fiscal audits provided by such corporations as Booz, Allen, and Hamilton. The audits are expected by both parties to resume annually until either party is no longer satisfied with the arrangement. Many of these auditing agencies have increased their staffing to offer program evaluation services. But here, too, there is little expectation that the persons working on the evaluation in a given year will have done so in the past and will build upon historical perspective. The format of the review is standardized to lessen the need for situational study. An interpretive council drawing from qualitative research methods would give particular attention to evolving situations.

The question may not be so much a matter of longevity of acquaintance as its intensity. Various corporations employ organizational and fiscal specialists to reside within the headquarters or plants for extended periods of time with a rather broad responsibility for discerning what is happening. When General Electric acquired the National Broadcasting Company in 1986, viewers were switching from the networks in great numbers to watch cable channels. Concerned about keeping the network profitable (Auletta, 1992) new Chief Operating Officer Robert Wright brought in a consulting team of four accountants to find ways of reorienting NBC away from revenue enhancement toward cost containment. GE officials wanted them to study "the culture" of the organization, which, through lengthy interviews, observations, as well as document review, they did. What the team provided were not indicators but hugely subjective estimates of what might be saved. They described the contributions of long-time NBC officials, especially those more bent upon providing public service than maximizing shareholder profit. The advice of the consultants was appreciated by corporate managers and disparaged by program staffs—but their interpretations were considered typical of what disciplined, intelligent observers will ascertain when they have sufficient opportunity to study a massively complex situation—not necessarily right but better than what was known before.

A long-staying internal but independent Council could be just as irrelevant as brief visitors and just as constrained as an internal team, but steps could be taken to increase relevance and minimize constraint. The Council could be guaranteed access, obligated by contract to raise critical questions, and insulated in various ways from intimidation. Such functions might be refined by the study of biographies of unique advisors such as Averill Harriman, Oscar Davis of the former U. S. Court of Claims, and Sam Messick of the Educational Testing Service. The Council could use its own internal workings to challenge observations and interpretations. In touch with principal investigators and evaluators, it could try out draft language and preliminary findings on program officers and other administrators. But mainly, it would serve as critical memory in the service of, but not dependent on, the science education program managers of NSF.

Drawing on the Qualitative Tradition- Whether or not an Interpretive Council is a good idea, the strategy of increasing the interpretive resources of the National Science Foundation should be considered. The present NSF investment in design of evaluation studies far outweighs its investment in interpretation. I have offered caveats here to recognize the

shortfall in efforts to build a rational evaluation enterprise. I have presented my argument here in terms of the epistemological flaws in evaluation data and indicators that might be used to define the effects of Foundation programming, claiming that the usual indicators of need, productivity, or systemic effect are largely hypothetical, created more from social theory and political discourse than from empirical science. These indicators belong to a largely fictitious world referred to here as virtual reality.

It is within the capability of the educational research community to improve the present battery of indicators, from the Wallchart on up, but the utility of indicators appears to be to enhance or justify decisions already made on political grounds (Lindblom and Cohen, 1979). Rather than develop and validate better indicators, as many qualitative and quantitative researchers would urge, my recommendation has been to increase the quality of interpretation available to program officers, central administration, advisory panels, and oversight committees. Much depends on peer review, and peer interpretation, not just those peers on a special council, but all Directorate members. According to Michael Scriven (1992):

> Like democracy, peer review may be a flawed system but, if given its best possible implementation, it's the best in sight and something like it will always be a key element in proposal and program evaluation.

The emphasis in this paper has been not on review of projects or proposals but on review of program performance. Such interpretive evaluation could be accomplished in various ways (with the 1978 advice of Cronbach and associates still highly pertinent) but probably not with major reliance on external contracting and rotatory personnel. Institutional restructuring is needed—bringing greater disciplined interpretation inside. That needed interpretation, comprehensive yet program-specific, would require enduring study under security enjoyed by judges and scientists. I think the most important contribution the qualitative paradigm can make to the evaluation of systemic effects is to raise the emphasis on disciplined interpretation.

Informal evaluation of systemic effects of NSF programs already takes place; more formal evaluation is said to be needed. These programs are part of a political process and their evaluation is part of that political process. Efforts to shelter the evaluation from political pressure are needed: they cannot expect to be entirely successful. The qualitative strategy of increasing personal interpretation responsibility in a formal evaluation effort requires long-term agreements and protection to those who will bring bad news. A pressure-free environment is unrealistic, and explanations by interpreters are another form of virtual reality. But validation, experiential as well as technological, can engage the merely virtual in improving understandings of program merit and worth.

References

Aaron, H.J. 1978. *Politics and the professors: The Great Society in perspective.* Washington, DC: Brookings Institution.

Auletta, K. 1992. *Three blind mice.* New York: Vintage Press.

Ball, D.L. 1992. *Implementing the NCTM standards: Hopes and hurdles.* Paper prepared for the National Center for Research on Teacher Learning, Michigan State University.

Barnouw, E. 1970. *The golden image*. Oxford, England: Oxford University Press.

Boring, E.G. 1950. *History of experimental psychology*. New York: Appleton-Century-Crofts.

Chelimsky, E. 1987. The politics of program evaluation. *Society* 25, no. 1 (November/December).

Chelimsky, E. 1991. The politics of dissemination on the Hill: What works and what doesn't. Paper presented at the Conference on Effective Dissemination of Clinical and Health Information, 22 September, at the University of Arizona.

Cook, T.D. 1978. Speaking for the data. *APA Monitor* 9, (3).

Cook, T.D., Cooper, H., Cordray, D.S., Hartman, H., Hedges, L.V., Light, R.J.; Louis, T.A.; and Mosteller, F. 1992. *Meta-analysis for explanation: A casebook*. Russell Sage Foundation.

Cronbach, L.J., et al. 1980. *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

Denzin, N.K. 1989. *The research act*. 3rd edition. Englewood Cliffs, NJ: Prentice-Hall.

Denzin, N.K., and Lincoln, Y. 1994. *Handbook of qualitative research*. Newbury Park, CA: Sage.

Eisner, E. 1979. *The educational imagination: On the design and evaluation of school programs*. New York: Macmillan.

Erickson, F. 1986. Qualitative methods in research on teaching. In *Handbook of Research on Teaching*, ed. Merlin C.Wittrock. New York: Macmillan.

Geertz, C. 1973. Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures*, ed. Clifford Geertz. New York: Basic Books.

Guba, E., and Lincoln, Y. 1981. *Effective evaluation*. San Francisco: Jossey-Bass.

Guba, E., and Lincoln, Y. 1982. Epistemological and methodological bases of naturalistic inquiry. *Educational Communications and Technology Journal* (Winter): 232-252.

Guiton, G., and Burstein, L. 1993. Indicators of curriculum and instruction. Paper presented at the AERA annual meeting, Atlanta.

Katzenmeyer, C. 1993. Addressing program evaluation in federal mathematics, science, engineering and technology education programs. Unpublished paper. National Science Foundation: Author.

Krueger, M. 1983. *Artificial reality*. New York: Addison-Wesley.

Lindblom, C.E., and Cohen, D.K. 1979. *Usable knowledge*. New York: Basic Books.

Miles, M.B., and Huberman, M.A. 1984. *Qualitative data analysis*. Newbury Park, CA: Sage.

National Council of Teachers of Mathematics. 1989. *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM

Polanyi, M. 1969. *Knowing and being: Essays by Michael Polanyi*. Chicago: University of Chicago Press.

Porter, A.C. 1993. School delivery standards. *Educational Researcher* 22, 5 (June-July): 24-30.

Psotka, J. 1993. An exploration of virtual reality. Paper presented at the AERA annual meeting, Atlanta.

Schwandt, T. 1994. Constructivist, interpretivist persuasions for human inquiry. In *Handbook of qualitative research*, eds. Norman K. Denzin and Yvonna S. Lincoln, Newbury Park, CA: Sage.

Scriven, M. 1967. The methodology of evaluation. In *Perspectives of Curriculum Evaluation*, edited by Robert E. Stake. AERA Monograph Series on Curriculum Evaluation, no. 1. Chicago: Rand McNally.

Scriven, M. 1978. Evaluating educational programs: The best models and their relation to testing. Paper presented at the Second National Conference on Testing, CTB/McGraw Hill, September, 21-22. San Francisco.

Scriven, M. 1991. *Evaluation thesaurus, 4th ed*. Newbury Park, CA: Sage.

Scriven, Michael. 1993. Hard-won lessons in program evaluation. *New directions for program evaluation*, 55. Summer. San Francisco: Jossey Bass.

Shavelson, R.J., McDonnell, L.M., Oakes, J., and Carey, N. 1987. *Indicator systems for monitoring mathematics and science education*. Santa Monica, CA: Rand Corporation.

Spiro, R. J., Vispoel, W.P., Schmitz J.G.; Samarapungavan, A., and Boerger, A. E. 1987. Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In *Executive control processes*, ed. B. C. Britton, Hillside, NJ: Erlbaum. 177-99.

Stake, R.E., ed. 1975. *Evaluating the arts in education: A responsive approach.* Columbus, OH: Charles Merrill.

Stake, R.E. 1986. *Quieting Reform*. Urbana, IL: University of Illinois Press.

Stake, R.E. 1994. Case studies. In *Handbook of Qualitative Research.* eds. Norman K. Denzin and Yvonna S. Lincoln. Newbury Park, CA: Sage.

Stake, R.E., and Trumbull, D. 1982. Naturalistic generalizations. *Review Journal of Philosophy and Social Science* 7(1-2): 1-12.

Strauss, A., and Corbin, J. 1990. *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

van Maanan, J. 1988. *Tales of the field: On writing ethnography*. Chicago: University of Chicago Press.

von Wright, G.H. 1971. *Explanation and understanding*. Ithaca, NY: Cornell University Press.

Woolley, B. 1992. *Virtual worlds*. Cambridge, England: Blackwell.

Yin, R.K. 1989. *Case study research: Design and methods*. Newbury Park, CA: Sage.

We've had several models of discussions this morning, and I am going to introduce you to a third model. I am also going to talk about two of the papers.

The papers I have been asked to discuss today are very different, as you have just seen. In one, Bob Stake looks broadly at the field of evaluation, notes its gaps and its failures, its distorted emphases, and its unresolved tensions, and tries to build an evaluation mechanism for NSF that could perhaps remedy these problems. Specifically, the paper speaks to the promise of qualitative research, to the needs for experiential understanding rather than explanation, for interpretation rather than a search for cause and effect, for the distinction of system patterns of information over time, and for the conciliation of historical perspective with independence (I guess you'd say "semi-independence," Bob. I noticed that changed in the evaluation function.) The proposal is for an invigoration of interpretive responsibility to be incarnated by a group of "semi-independent" evaluation researchers within NSF. The group members would do some evaluations, advise on others, and generally lend their research expertise to the improvement of agency evaluation information over time.

The second paper describes a particular method—cluster evaluation—and proposes it as one likely to be useful to NSF in addressing two needs that its authors, Zoe Barley and Mark Jenness, judge important in the evaluation field today: the need to account for and conciliate the use of stakeholders, and the need to structure evaluations to serve the primary function of improving the program.

So, one paper focuses on a particular evaluation method, the other on a broad approach to assessment. One emphasizes knowledge, the other stresses the program and its services, but both papers deemphasize the importance of attribution of defined outcomes. I read both papers with great pleasure and think them worthy of NSF's careful attention and reflection.

Cluster evaluation seems to me to be a reasonable way of achieving buy-in and consensus across what are often warring groups. It's less clear to me how findings could be developed from the analysis —again Bob Stake's point about the need for validation—and whether so complex a process would be both feasible and productive.

Bob Stake's paper, which is a sort of luminous meditation on the problems and joys of producing something like real knowledge through evaluation, brings some critical insights to the assessment of teaching and learning. Reading his discussion of the distinctions between quantitative and qualitative representations of realities, I was reminded of the passage in Gabriel Garcia Marquez's *100 Years of Solitude*, in which the town of Macondo loses its memory and is forced to put up signs reminding citizens of the names of objects and how to perform functions like milking cows. By the way, the first object for which a sign is made is called a stake, spelled S-T-A-K-E, and of course another sign tells people exactly how to go about milking cows.

It's true that signs and other "virtual" quantitative abbreviations cannot represent everything, but sometimes it's the best we can hope for. My own bias in looking at an evaluation function—that is, how it should be organized and what methods are most valuable—would add some other components to those presented in these two papers. To me, the kinds of evaluations that need to be done will always depend heavily on three things: the kinds of policy questions or evaluation questions that will be asked about the program, the service, or the function; who will be asking these questions; and what evidence will be needed both to answer the questions and satisfy the political and institutional culture of the particular audience. The question, after all, is the critical trigger that determines what methods need to be used.

Someone asked the question earlier, Can we really separate evaluation from dissemination?

Again, that depends on the question. If we are looking at something that the Congress might ask us to do—for example, evaluate a study and tell us whether it's good—we would simply do an evaluation of it. We would critique it in one way or another, depending on what the study was, but there would be no need for dissemination other than simply passing it to the committee that wanted it. If we are talking about a program where the question is, Can we use intermediaries to disseminate knowledge to a given audience? then dissemination is part of the evaluation—it can't be separated. So it all depends upon the question that is asked.

I think we shouldn't forget that traditional quantitative and qualitative methods can answer a great many questions about the effectiveness of programs or functions and the quality of services (for example, questions about whether someone learned something or not, or whether program beneficiaries are pleased with or insulted by the services they receive). But ingenuity and creativity and innovation are needed to answer broader, complex, systemic questions.

To me this suggests four interdependent means of dealing with these broader issues. The first is an evaluation organization that starts with a profound understanding of which questions will most often emerge, and why, from the political environment within and surrounding an agency and its programs. The second is a panoply of traditional methods and the skills to apply them appropriately and to validate them. The third is the exploration of new methods as a response to questions that cannot be answered with old ones, and the fourth, an in-house organization that can demonstrate the feasibility and usefulness of doing both the old and the new. New methods cost a lot of time and money to specify, test, and apply, and they involve some risk to their users. In particular, the more political controversy there is about a topic, the greater the initial credibility risks of newly developed methods. Therefore, the evaluative requirement for them should be, I believe, abundantly clear and their use warranted by the need for answers to specific user questions.

I want to begin by expressing my gratitude to Zoe Barley, Mark Jenness, and Robert Stake. I want to thank them for giving me the opportunity to read their papers and learn from them.

My thoughts are organized into four themes. First, ideas, solutions, and innovations have difficulty moving horizontally in hierarchical systems. Second, local-level project personnel in social programs can do program evaluation, if technical assistance is available. Third, qualitative analysis is central to the evaluation process. And fourth, NSF needs to study the problems of math and science education in a larger social context.

Promoting Horizontal Flow of Information

I have a lot of experience working in local-level programs, and I have learned that information usually flows vertically in any institutional system. Reports, plans, audits, monitoring results, evaluations—all of this stuff moves from program units through management to policy people. Few resources are given to moving information between program units. Consequently, the people who are responsible for delivering services in a program rarely have means or opportunity to communicate with each other.

Cluster evaluation, as described by Zoe Barley and Mark Jenness, does much to overcome the horizontal flow problem. In the cluster approach, regular networking conferences for the projects are a central feature. Staff from different projects participate in negotiating agreed-upon common outcomes and then collaborate in data collection. Finally, "dissemination of findings and sharing of lessons learned occurs between individual projects in the cluster..."

Local-Level Evaluation

In my current job as Director of Program Evaluation for ACTION, the Federal domestic volunteer agency, I been actively engaged with the problem of how to get project staff involved in evaluation. My agency gives grants to community-based organizations. Many of those grants carry a congressionally mandated requirement that they conduct an annual evaluation of their programs. For small grants, say under $100,000, this may appear to be an absurd requirement. The resources needed to meet the evaluation standards of the grant guidelines are seen by project personnel as detracting from their basic mission, which is not research. In small programs, often the evaluation tail is wagging the service delivery dog.

Through ACTION training conferences for grantees, I have made some efforts to overcome this problem. I try to give project personnel some skills in what I call local-level, low-tech, low-cost evaluation techniques. For example, I ask participants (and sometimes I might have a few hundred in a room with me at one time), "How many of you know your annual budgets?" Everybody raises a hand. Next, I ask, "How many of you know how many hours of volunteer service your project produces each year?" Almost everyone raises a hand. Finally, "How many of you calculate the cost per volunteer hour of service?" Rarely have more than 3 or 4 persons in 100 responded affirmatively.

Again, cluster evaluation proponents recognize this problem and opportunity. The cluster evaluation approach emphasizes the central involvement of evaluation in program management and improvement and stresses the importance of direct stakeholder involvement in that evaluation. The processes of cluster evaluation, as described by Barley and Jenness, go a long way toward empowering local-level project people with needed evaluation skills and other resources.

Qualitative Analysis

In reading Bob Stake's paper, I was reminded of a time years ago when I was doing extended field

research in Johnson County, Kentucky, the birth place of Loretta Lynn. In some of the Pentecostal churches in that part of eastern Kentucky, there was the belief that a person possessed by the Devil could not say the word, "J-, J-, J-,Jesus!" Well, Bob Stake apparently is possessed by some demon for he cannot say the word "A-, A-, A-, Anthropology!"

He refers to several concepts and methods that are the traditional domain of cultural anthropologists. These include ethnographic research, the emic/etic distinction, and holism. In one passage, he presents a fair representation of anthropology's central concept, culture.

We have common knowledge not because there are pre-existing facts—truths—for us to discover, but because learning, just like dressing and driving, is a social process. We have strong tendencies to conform. We modify our actions to fit the actions of those we respect. And we create knowledge very similar to that of the people around us.

Stake mentions several of the social sciences, but nary a mention of the father and mother of qualitative research, anthropology.

I recommend to this audience the extensive research literature in applied anthropology. In this subdiscipline of anthropology, the concepts and methods that Bob Stake discusses are not nontraditional, rather they are very central to our tradition.

One caution: qualitative research is not easy. Bob Stake is absolutely right in characterizing it as costly, time consuming, and subjective. My experience with contractors conducting research for my agency may be typical. Our research designs often call for site visits, case studies, and other types of participant observation. I have yet to see the wealth of information gained in these qualitative approaches used in any way other than as anecdotes to fill out quantitative reports.

I would disagree, however, with his contention that the "results pay off little in the advancement of social practice." While a reply would need another paper, I must say that applied anthropology has made major contributions to improving social conditions, especially in the developing world. One example is the important role that anthropological (qualitative) research is playing in the development of techniques to disseminate health information on AIDS in Africa.

The Larger Social Content of Math and Science Education

As a final comment, I want to suggest to the National Science Foundation that it expand its research on the problems with math and science education in the United States. In addition to improvements that might be made to curricula, we need more understanding of the cultural settings for science education in our country.

While we are a nation that seems to revel in technological advances, we are also a nation beset with rampant superstition, ignorance, and even rejection of basic scientific processes, principles, and theories. Almost a majority of people in this country, if some recent polls are to be believed, accept the creationist view of our origins (the story in Genesis) and reject basic evolutionary theory. Millions profess to believe in astrology. The list of irrational belief systems that are being embraced by substantial numbers of Americans is quite lengthy.

The question for NSF is, How can we educate children in science when their parents show such disregard for its most basic principles?
