Overview

Michael Scriven

I want to begin by saying how important I think meetings like this are, that is, meetings in which the existing paradigms of evaluation are seriously questioned by those who are not only involved in the game, but also those who are hiring these people and those who are being evaluated by these people. I think we should regard it as a kind of moral imperative for evaluation as a discipline that meetings like this happen.

The results of the major efforts that we have heard about today are impressive. One of the results is a series of suggestions on a very practical level, in particular, a list of 40 suggested questions that you might ask in doing an evaluation of programs like the examples from NSF. There is no substitute for the local experience that some of these people have as evaluators and as program participants. While their comments are aimed at NSF many of them will work equally well for another agency. Many are generic types of questions, though specific enough to be relevant to the ground level of evaluation. So, I think simply on that ground alone, we have something worthwhile here.

On the other hand, there was, I thought, a substantial lack of clarity about what was being done in the efforts discussed today. That doesn't mean that they're not useful. It's just that the interpretations given them were sometimes implausible.

The three things that were going on in these papers, apart from trying to improve evaluation, were:

- Trying to improve dissemination;
- Trying to improve explanation and understanding; and
- Trying to improve description of process—what happened? How did it come about?

These three things need to be distinguished, not sharply—that's not possible—but generally speaking, as carefully distinguished as possible. I think we are meant to be talking about evaluation. Let me put it another way. Dissemination is a specific process that's crucial in certain projects, but absolutely irrelevant in others (e.g., where you are trying to solve a theoretical problem, and the payoff is having solved it). The justification for the project is that it had a reasonable chance of solving the problem, not that it did solve it. Dissemination, as Eleanor Chelimsky put it, is going to come in if the task of the evaluation is to find out whether the results were disseminated successfully, and it's not going to come in if the task of the evaluation was to find out whether the problem had been solved, useful discoveries had been made, etc. I think this distinction is quite unclear.

"The results of the major efforts that we have heard about today are impressive."

One of the reasons for that lack of understanding leads to a constructive conclusion that we should take extremely seriously. We really are not treating dissemination as a research area, although it's very unfortunate that we are not. We're constantly reinventing wheels, or much worse, we're starting to realize that someone already did, but we don't know how. There are lots of tricks out there in "dissemination land," and even some experts in some parts of it, as you well know. But we're not treating it as a body of knowledge we must have to get many of our tasks completed.

Dissemination is, of course, a perfectly sensible part of applied social science. We just need to give it more attention and expect to get more from it. Then, we can pull that knowledge in without having to force people who want to help in changing the schools to be experts on dissemination, which many of them are not.

"I do not have the faintest understanding, nor does anybody else, of why aspirin works. But as an evaluator in the pharmacological field, it's not a big problem to prove that it does."

The explanation and understanding issue is a little trickier because there is a gray area. Bob Stake spent quite a bit of time talking about the importance of qualitative research as a way to obtain insight and understanding, perhaps on the way to explanations of certain kinds. Well, there is a part of evaluation where explanation and understanding is of the essence. It's what you might call "perspectival" evaluation, where what you are doing is trying to achieve a new perspective on the program — to see it in a different way. Wittgenstein spent years toward the end of his life working on the phenomenon of seeing one thing as another thing. That's a very important part of what the good evaluator, and particularly a good qualitative evaluator, can do. But it's only part of the job, and it's only part of the job in some kinds of evaluation tasks. So, we want to be careful about thinking that explanation and understanding is, in general, part of the evaluation job. It is not. I do not have the faintest understanding, nor does anybody else, of why aspirin works. But as an evaluator in the pharmacological field, it's not a big problem to prove that it does. I don't want to be fooling around too long with people who keep saying, if you can't understand how learning goes on, then you can't evaluate teaching. Of course I can evaluate teaching; I don't need to know anything about learning theory, I just need to be able to recognize effective teaching when it bites me.

So we don't want to get into this academic trip about the need for the theories in order to do good evaluation. On the contrary, in many cases, if you can't do good evaluation, you can't even develop the theories of good teaching. Evaluation is the groundwork without which you cannot validate the theory. You want to know what methods of teaching work better, so you need to have measures of learning, not the theories of learning, which you can use to find out which methods did work better. You must be able to evaluate the learning, assess the students' work in order to evaluate the theories.

Indeed, there was one paper which was almost entirely devoted to discussions of questions about how things happened (i.e., descriptive research on various processes in learning and teaching). That's important stuff, it is a part of the task of the RTL program, but it's not a part of the task of evaluating teaching.

There were thus four kinds of valuable payoffs from the papers and comments. First, we were presented with a wonderful array of suggestions for indicators and questions to be asked when evaluating important programs of this general type. Second, there were suggestions for needed research on evaluation. Third, quite a different matter, there were suggestions for research on how teaching and learning works. And some of the suggestions for research on evaluation were, in fact, suggestions for research on dissemination or research on explanation. We can shuffle those over to other groups where they are useful topics for research, but not of direct concern for use in evaluation here.

Fourth, there are the proposed "new models," and one aim of the conference was in terms of looking for new models or approaches. Here I think the arguments are less persuasive, and I find myself in the truly embarrassing position of defending the status quo, something which I've never done throughout my life. But there doesn't seem to be anybody else around to say, "Hey, that's a straw man, we do better than that today." So I'm going to argue in that direction for a while.

We need to distinguish first between the arguments that we do need a new approach, and specific arguments for the proposed new approaches. We have heard quite a few of both of these. The arguments for needing a new approach are, in my view, mostly aimed at what is really a straw man. Now, NSF has had a great deal of experience with the standard approach to evaluation because it sends out a lot of RFPs to get evaluations done and it sees what comes in. So I'm not going to second guess their view, that there's a body of bidders who trot out their favorite quantitative something or other model. Yes, things creak at the joints a bit in the process of development, but one doesn't really want to treat that as the state of the art. If we're going to start looking for new paradigms, then we need to see if the existing best practice is faulty. And the best practice isn't always what Brand X trots out with their number 16 proposal writer when you run an RFP up the flag pole. Best you can get from Eleanor Chelimsky; the best you can get from the best of the audit agencies; practice is the best you can get from the best of the OIGs; the best you can get from the best practitioners in the American Evaluation Association none of whom are bidding on these RFPs. We want to be careful that we don't rush to ditch current best practice on the grounds that current proposals are unsatisfactory for the sort of tasks that are involved in evaluating the types of programs exemplified, but not restricted, to the three big NSF programs that were mentioned frequently.

It seems to me, for example, that the best current practice is a kind of eclectic amalgam of qualitative and quantitative. It's certainly not just quantitative. And this is not only for the reasons Bob Stake gives that there is no such thing as pure quantitative, but also for the other reason that these days best practice will have explicit qualitative elements aimed at various areas such as those where you can't get a good quantitative grip and those where the interpretive process is absolutely fundamental. Numbers aren't going to do the interpretation for you. So, it's an eclectic mix of quantitative and qualitative, formative and summative, internal and external, worth and merit. That is, it involves looking at cost effectiveness and not just effectiveness.

"We need to distinguish first between the arguments that we do need a new approach, and specific arguments for the proposed new approaches."

In the *Call to Arms*, Joy listed the reasons that the Directorate had for suspecting that there might be a need for a new paradigm. She says that the traditional approaches are "not directly applicable to the many research-oriented, ground-breaking, inquiries" that NSF often supports. Well, of course, "ground-breaking" is an interesting phrase; it does suggest that you broke some ground. And, if you broke some ground, it does suggest that there ought to be some sort of a footprint in the sand. We should at least see some sort of a new path, some blockage that got broken through, some problem in conceptual understanding was solved. So I don't feel that we really should have to say, "Abandon hope all ye who enter the eclectic, contemporary model of evaluation," here's a case where you can't handle the challenge. Groundbreaking is easy; at any rate groundbreaking is a lot easier than, "Did it have an effect on the kids in the 12th grade in the United States?"

The research efforts in RTL, for example, are in an important sense, much easier to evaluate in themselves. However, the question of whether everybody has come to recognize the leading work in the field, whether the practitioners have all been affected by these efforts, is the dissemination question. It involves another step, and it's harder. The question of whether the problem is solved is not so hard. And so, I think it's a really serious reason for avoiding naive applications of a quantitative model, which you certainly run a risk of getting when you put out RFPs. But you should expect to write your RFPs to rule out the naive bidders, expect to be very tough about awarding contracts, and restrict awards to people who see through the simple-minded ways of handling the issue at hand.

Joy adds that the impacts are different between studies that are research oriented and those that are groundbreaking. For example, she says that the old style of ground-breaking evaluation "seeks to attribute the effect to a single source." Well, is that really true? They were interested in the question of whether somebody's project did it, if that is a single source, because that's what they were asked to find out. But, then you can hardly blame them because they looked at the question of whether a single project did it. I find myself wading through many pages of their variance analysis, which says, 'No, there isn't a single source that did it, but the single source contributed something to it, here's the figures to prove it.' That doesn't seem totally stupid to me; it seems to me that's a fairly sensible kind of approach. So I think we can handle the notion of more than one source, and even the quantitative fellows, bless them, actually do that quite a bit, and certainly the rest of us can do it too.

The second thing she says and, of course, Joy didn't invent all this out of whole cloth—she's picking up common comments—is that standard evaluations are almost entirely reliant on quantitative data. Well, that is a sign of weakness in the bidder, in the evaluator. Let's not make any mistakes about it, if they're almost entirely reliant on quantitative, then in very many cases that will be just a flaw in their capacity to solve the problem of getting a true measure of merit and worth. But that seems to me to be an example of bad use of a simple-minded paradigm, not an example of current best practice being unable to handle the problem.

Following up on this point, she says that quantitative won't do because a single successful project may justify the entire research investment. Indeed, but where do we have somebody saying the program was a failure because only one Einstein went through? Nobody says that; or if they do, then scrape them off the list for next time around.

We can cope with selecting portfolios of high-risk, high pay-off investments. At the first meeting of the Evaluation Network, 20 something years ago, I set that task as the task for the President's prize. Nick Smith won the prize for a study in which he showed how to handle portfolio assessment. It's discussed in some references as the apportionment problem. So, we want to be careful about hopping on a new bandwagon on that issue. I'm speaking reluctantly in favor of the existing best practices being better than you might think.

One of the things I do at the moment is handle all the external evaluations for a wealthy community foundation that funds absolutely everything you can think of—legal aid, work in San Quentin, housing for dispossessed mothers, help for the drug addicts, restructuring schools. Mention anything, we've got a program, probably six. Now, that's a very wide variety, but we don"t find any need to shift paradigms among them. In fact, the value of somebody handling a wide variety of evaluations for the trustees of the foundation is that they can use a consistent model across the board. It gives them a degree of comparability which is useful. Perhaps we ought to think the same way about large agencies. We should be trying to use a standardized model—which doesn't mean a primarily quantitative model—across the board.

Then there was the question of the tendency to give priority to measures of student achievement. Well, is it an inadequate sole measure for some NSF programs? Certainly it is, and if you were to use that as the only measure, you would have to wait around 25 years to get some of the data, which wouldn't be much good.

So the real rival for the new style religion is the reformed orthodox church, not the church of the 1960's. Bearing that in mind, we now come to look at the proposed new models. These are not very much of a threat to the reformed orthodox model; they are much better seen as suggestions which should be used to forge refinements of the eclectic best practice model. I think that they can be very useful in that role. Cluster evaluation for example, seems to me an excellent device for improving evaluation, if we redefine it. Redefined, it looks something like this. The evaluation staff, on a group of related projects, regularly meet to discuss what they are doing and how things are going; and occasionally, but only occasionally, meet with the project directors in order to discuss how things are looking, but in limited terms, not full disclosure at all. In the way in which this was described to us here, it was really a replay of the original, transactional, North Dakota, East Anglia, model of collaborative, negotiated evaluation. Which, to put it bluntly, is a great way to cheat the consumer. Who's represented at the negotiations? It's an exact analogy to the way in which the union meets with school district management to thrash out the contract. Who's not there? There's nobody representing the kids, nobody representing the taxpayer. And you get just the same amount of credibility with the results. So, in this case, getting the project people in bed with the evaluators is exactly what you do not want to do if you want a credible and serious evaluation. Now, that approach is very popular these days; the President of the AEA calls it "empowerment evaluation." But it's simply a way to guarantee the loss of what objectivity is possible in those ongoing, formative evaluations, and that's a terrible loss. Why do you read Consumer Reports? Why don't you just read the handouts from General Motors? Well, suppose we insisted that the Consumer Reports auto evaluators spend the year with GM engineers. Will that improve the objectivity? No, it will corrupt it. We knew that from day one. So, I don't feel happy about that example.

"... getting the project people in bed with the evaluators is exactly what you do not want to do if you want a credible and serious evaluation."

It seems a bit mean to have picked on the cluster evaluation protagonists and then not to go pick on everybody else, which I could easily do. But instead, I'm just going to do two remaining things. First, I'm going to put forward what Bob Stake will regard as a truly straightforward demonstration of my simplemindedness, by defending the silver bullet approach. Then I'm going to talk about Bob Stake's paper.

Now, I'm going to ask you in thinking about this intervening discussion where I want to convey to you, what I believe we ought to be doing, to think of three people. The first is Mosteller, whose name was mentioned earlier. Fred Mosteller at Harvard is generally thought to be one of the two or three best applied statisticians in the world. He's the author of *Understanding Robust and Exploratory Data* which was a reality-oriented push in statistics. He is also the author of another notion which I want to commend to you today because I intend to use it as a paradigm. After years of editing a journal and receiving countless submissions in which something or other turned out to be statistically significant at the .05 or the .01 level, he coined the term, "interocular differences" to contrast with "statistically significant differences." His line about them is very simple. Go ahead and play around with the statistically significant differences while you are doing research because it may help you find something interesting. But don't come to me until you've found some interocular differences. In other words, if the difference doesn't hit me between the eyes, I don't want to hear about statistical significance. Now that's the voice of a good statistician and it's a very sensible appropriate voice when you look at what happens to the 95 percent of published research that was statistically significant. It doesn't replicate the second time around, it turns out to be trivial in the light of various conditional requirements on it, and so on, and so on, and so on. So the first point is, we ought to be looking for interocular differences in evaluation and we ought to be sending the statistically significant stuff back to the drawing board.

Now, the second person I'd like you to keep in mind, though you haven't ever heard of him, is John Hattie. You'll hear a lot about John Hattie. He's a brilliant eclectic educational researcher, my fellow professor at the University of Western Australia for several years. He's done an analysis of the kind that will make Bob Stake want to bring his lunch back, the kind of study which Congress just loves to get. It's this. He's looked at every educational intervention that can be given a generic description, such as should we add paraprofessionals; should we put computers in the classroom, in what ratio; should we reduce class size; should we increase inservice education; should we mainstream; should we ability group. He simply lists them, and does a meta-analysis, or finds another meta-analysis that has already been done on each of them. He finds the effect size and lines it up, and he says, if you've got X bucks you can possibly spend in a school district, here's the shopping list in order, this is what you'll get for each buck.

You'll remember that Hank Levin has done a very nice study of that kind, aimed particularly at whether you should computerize or not but covering other things. Hattie has a generalized version of that. Of course, this will not be a perfect guide, but as Bob Stake says, we have to move from initially misleading indicators to better indicators. Now that's the kind of result that Congress is always pounding us for and that academics sneer at, but I think quite wrongly. In this connection, one should remember the story of the Office of Inspector General. There was one Inspector General 15 years ago, and there are 26 today. Why? Because the academics would never get the evaluation reports in until long after the people who needed them had left. An Inspector General finally said, I think it can be done in 3 months for $100,000, and so let's see. And, so now we have a whole bunch of people doing those evaluations. Have the academics ever done an evaluation study to show that these are such trashy results that they have led to millions upon millions of wasted money? No, they have not. Now that either shows that they don't want to find out, or that the results aren't at least obviously disastrous. So, I think exactly the same thing applies here: meta-analyses should guide policy. We want to be very careful to try to speak the language of common sense on these things.

I'll bring that down to cases. In the Advanced Technology program there is a great deal going on, but in 25 years of serious work in the Ed Tech area, I have found the same problem to be endemic that I see in the material here, briefly described though it is. You might sum it up by saying that they'll never look at the top competition. If you're looking for magic bullets in the Ed Tech arena, you won't find them by test firing against bows and arrows. Magic bullets have got to be the ones that beat the best of the other bullets; it's not interesting that they can beat bows and arrows. And we're finding a lot of material here whose only claim to fame is that it can beat a bow and arrow.

Specifically, there's very little in Ed Tech that can beat a programmed text, but we never run things in Ed Tech against programmed text. We run them against the status quo, non-Ed Tech approach, or against very primitive Ed Tech approaches. That's not serious evaluation. Programmed texts have now gone: "everybody knows" that they don't work. But there were many out there that could beat anything. They could beat intensive tutoring, they could beat the best teacher there was, they could beat what existed then in the way of computer-assisted material. And, so we just walk past that; we averaged it out. Who cares about the average? The question is, what was the state of the art? Certainly programmed texts were more expensive than standard texts, but a lot less expensive than most Ed Tech. So, one of the problems that we've got, is that the group of Ed Tech folk, are, to put it bluntly, massively biased in judging proposals. What is the effect on them of using the toughest possible standard, competing against the best alternative there is? It is that very few of them will ever be funded. They know that very well, so that you must understand that a lot of what I have to say consists in saying, don't do collegial review, don't talk peer review, if by that you intend to mean people from the same in-group, because they are massively biased.

Now, with respect to Bob Stake's final suggestion about a panel, I'll suggest how one might expand that notion, so that you would, in fact, get quite a good degree of independence. When you do a secondary school accreditation, it's always a bad deal because when the team of 40 arrive at the high school, it's got one person on it in Driver's Ed, and one person in Accounting, and one person in whatever, and after the Driver Ed person goes to look at Driver Ed and has tea with his friends he saw last week at the All-State Conference in Driver Ed, he then comes back saying, "Gee, this school is strong in Driver Ed." What's that worth? Nothing. If you'd sent the accountant to look at Driver Ed and the

"There was one Inspector General 15 years ago, and there are 26 today. Why?"

Driver Ed guy to look at Accounting, we might have learned something. Better, send both to both. We should use that model for panel construction—the mix of local and outsider expertise.

So, remember Mosteller, remember Hank Levin on the employment futures that high tech delivers and on the relative payoff of various ways you can spend money on student outcomes. Remember John Hattie doing that more generally, and me talking about the programmed texts as the main competitor with CAI, e.g., with enormously expensive PLATO installations. I did the largest evaluation of a PLATO installation that's been done so far, so I speak with some interest in that area.

The bottom line of that sort of study, from Mosteller through Hattie, is the sort of thing that Congress rightly wants to see. Academic condescension says, 'No, that's a naive assumption about how easily you can produce indicators for these things.' I think not. I think the fact is, that we ought to revitalize the entire effort so that the task is this: using the Ed Tech area as an example we'll give you a little money for a pilot; then if you show signs that you can beat a programmed text, we'll re-fund you for a limited period of time. If we want magic bullets, we have to set the shooting competition up with the proper rules; beat the best, or go back to the drawing board.

"If we want magic bullets, we have to set the shooting competition up with the proper rules; beat the best, or go back to the drawing board."

Footprints: A Search For New Strategies For Evaluating EHR Programs

Laure Sharp and Joy Frechtling
Westat

Prologue

This paper presents our interpretation of what was said at the "Footprints" conference and written in the "Footprints" papers. It is not an attempt to summarize all suggestions or to comprehensively discuss the pros and cons of each author's proffered strategies. Rather, we have attempted to extract the points that we see as especially relevant to the Division of Research, Evaluation and Dissemination (RED) and to offer our suggestions for how RED can build on what was learned from the "Footprints" task to shape its future evaluation agenda.

Introduction

In 1994 and 1995, several programs funded by NSF's Directorate for Education and Human Resources (EHR) are scheduled to undergo third-party evaluations. Planning these evaluations will be a complex task, given the heterogeneous nature of the programs and the projects that they support. As a first step in the planning process, the National Science Foundation asked Westat to commission a series of papers from experts in diverse fields of evaluation to help develop a framework for examining these programs. The eight commissioned papers and the comments of seven discussants are presented in this volume. In this final paper, we have sought to highlight and discuss those topics and ideas that emerged from the conference and seemed most germane to EHR's planning needs. This selective review was guided by what we believe are EHR's concerns and especially those of RED in undertaking program evaluation in the near future. Many more valuable ideas and comments can be found in the papers and discussions, and they deserve close review by NSF staff and others interested in innovative evaluation practices.

The Need for a New Evaluation Approach

New techniques were sought because the RED staff felt that traditional educational evaluation methodologies would not be appropriate to assess what many EHR programs had accomplished.

Traditional evaluations of educational programs have been developed primarily to assess the results of new or improved service delivery models. For example, Chapter 1 and Headstart typify the service delivery model and provide the template against which most large scale federal evaluations have been constructed. In such evaluations, typical questions include the following:

- Do students benefit from the introduction of new services or technological innovations, such as the use of computers?
- Do students' attitudes, interests or test scores change?
- Do teachers adopt new instructional methods after attending science workshops?
- Do these new methods result in improved student performance?

The service delivery model may be appropriate for some EHR-funded projects. However, it is ill-suited to many others, and with a few possible exceptions, it is inappropriate for the evaluation of programs. The mismatch stems from a number of sources, including the organization and makeup of the EHR programs, the goals the programs are intended to meet, and the very nature of the funding mechanism that predominates.

Each of these is considered further below.

Program Structure

Traditional evaluations have been developed to assess the impact of programs supporting projects that are fairly homogeneous in nature. They have common components and may even be built along a "planned variations" model. EHR programs, including Research on Teaching and Learning (RTL), Applications of Advanced Technologies (AAT), Studies and Indicators, in contrast, support a wide variety of projects that are highly diverse and vary in size and duration. Some are part of a stream of research, reflecting decisions made over multiple funding cycles. Some reflect the results of cross-program collaboration. Others are one-time efforts or exploratory projects.

While some of these projects can be evaluated using a service delivery model, for many others the model is unsuitable or, at best, incomplete. For one thing, it cannot be applied to projects that can be categorized as basic, theory-driven research (as contrasted with those categorized as applied, problem-based research). It is also inapplicable to descriptive studies and those that are funded by the Studies program to generate new international statistics on student achievement in mathematics and science (SIMS and TIMSS).

Even where the model may be applicable to individual projects, it is rarely appropriate for the evaluation of a program as a whole. That is, in many cases, it may be neither possible nor conceptually correct to aggregate individual project evaluations for the purpose of evaluating the program as a whole, if only because a comprehensive program evaluation must answer questions that go beyond assessing the outcome of individual projects. For example, to evaluate the RTL program, policymakers and other stakeholders may want to know if the funded projects addressed the most important research questions or had an impact on classroom practices in school systems other than those in the project sites. Aggregating the evaluations of individual projects does not provide answers to these more global questions. Furthermore, some programs - of which AAT is the prime example - may choose a "high risk - high gain" investment strategy, anticipating that only a few projects will lead to scientific breakthroughs. In this case, an evaluation based on aggregation of project outcomes would be especially inappropriate.

Program Goals

A second obstacle to using the traditional, service delivery model for many EHR programs is their broad-based and highly ambitious goal structure. Traditional evaluations have frequently been motivated by, and structured to address, specific legislative mandates. Rightly or wrongly evaluators have relied primarily on narrow goal specification and looked for indicators that can document goal attainment over a period of a year or two or even five.

The EHR programs on which we are focusing lack specific, tangible goals that are to be met within a given time period. While the ultimate objectives of NSF's programs in education and human services are clear, they are also very ambitious and very broad. The programs serve to promote more participation and better learning outcomes in mathematics and science among students at all educational levels and/or more recruitment into scientific careers especially for underrepresented populations. It is very difficult to assess progress toward these goals in the short time span under which program evaluations must typically operate. Further, given the magnitude of the implied task of changing major components of the educational system, holding the relatively modest NSF programs accountable for their attainment is unrealistic.

The Funding Mechanism

Perhaps the greatest obstacle to the use of traditional evaluation strategies for NSF programs stems from a third cause —the funding mechanism. Educational programs and projects for which traditional evaluations have been carried out were usually funded through contracts or grants that prescribed performance requirements, benchmarks, and outcome criteria. In the great majority of cases, EHR programs are based on the academic grant model, where grants are awarded to field-initiated projects selected through peer review. In this process the emphasis is on quality of performance and the qualifications of the principal investigator. Awards based on the academic model encourage experimentation with innovative ideas and processes; the grantor will, therefore, accept a high risk of failure as part of the research design. The process is tolerant of considerable deviation from proposed activities in the detailed execution of the project, at the discretion of the principal investigator, and gives investigators considerable leeway in their choice of procedures; adherence to specific performance criteria is seldom required. This grant model is in line with NSF's basic funding mechanism and philosophy for the support of research in the physical sciences.

As a rule, institutions using the grant mechanism to fund projects do not carry out systematic program evaluations. Rather, grant programs sponsored by government agencies and private foundations have relied for evaluation on judgmental approaches through expert panels, review committees, and similar mechanisms. Education programs are also being reviewed in this manner, but the mandated periodic third-party evaluations call for more systematic approaches.

Thus, RED must develop a strategy for the systematic evaluation of EHR programs whose goals and funding mechanism often preclude the use of methodologies traditionally used in the evaluation of education programs.

The Guiding Concept Proposed by NSF: Footprints

Understanding the difficulty posed by the need to evaluate many of EHR's programs, NSF staff sought new ways of examining program accomplishments. The "Footprints" model was chosen because it seemed to offer a new way of thinking about results and because it seemed flexible enough to apply to the evaluation of the very diverse programs funded in EHR.

"Footprints" were defined as evidence that the program had left a mark on the field of mathematics and science education and had contributed to new knowledge or new practices. Specifically, this metaphor suggests that the program evaluation should seek to ascertain whether a program has contributed substantially to the state of knowledge in mathematics and science education (the "research base"), and has left its own "footprints in the sand" (evidence that both researchers and practitioners have been exposed to this knowledge and/or have been influenced by it). A footprint implies that a mark has been left, but it is not explicit with regard to how and when the mark actually got there. This metaphor has the advantage of not being overly specific as NSF's Susan Gross said in her introductory comments, "Footprints come in all sizes and shapes," thus avoiding a priori restrictions on potential outcome indicators. RED staff initially identified four general areas where footprints might be found:

- Effects on the profession (the supply and characteristics of researchers, topics presented at conferences, and in journal articles);
- Effects on other research;
- Effects on practice (teacher training, curricula, and implementation of sound pedagogy); and
- Effects on funding agendas of other institutions.

Such footprints might begin to answer the broader questions which NSF itself, as well as oversight agencies within the Federal Government and congressional bodies, ask about these programs:

- What has been their impact on the thinking and practices of educators and administrators in local school systems?
- Are these programs likely to contribute to the achievement of national goals such as higher participation by women and minorities in mathematics and science education?
- Is there any evidence that they have improved the quality of instruction in science and mathematics at various levels of the educational system? Have the programs affected the thinking and actions of educational policymakers, of researchers, and of those who fund research at the national, state or local levels?

Ideas and Suggestions from the Conference Papers

As might have been expected, given the diversity in their backgrounds, work settings, and disciplinary orientations, each paper author and discussant came with his or her own experiences, approach, and ideas. While some presenters dealt extensively with the "Footprints" theme, others addressed the issue of nontraditional analytic techniques or, more broadly, the topic of nontraditional approaches to educational evaluation. As Joy Frechtling pointed out in her introduction to the conference, while none of the papers went so far as to propose a specific evaluation design for one or more EHR programs, they provide valuable directions and inputs. Many of these can provide useful guideposts as RED undertakes its planning efforts for third-party evaluations of EHR programs.

As we have thought about what was learned from the "Footprints" effort and attempted to distill the main points from what was said in the papers, by the discussants, and by the general audience, we have identified two "messages."

- Message 1: There are a number of alternatives to the service delivery model that might be applied to EHR evaluations. Indeed, what we have referred to as the traditional model may be traditional in only a very limited context.

- Message 2: There are many different frameworks that can be used to evaluate EHR programs on which we have been focusing. The footprints we have started to uncover lead in many different directions. Before choosing a direction for any specific evaluation, the audiences for the evaluation and their general interests/concerns must be defined by EHR.

In the subsections that follow, we discuss these messages in somewhat greater detail. Specifically, we will examine the following topics:

- Who is the audience for EHR evaluations?

- Is there a set of core topics that all evaluations should address?
- What techniques are suitable for proposed evaluation tasks?

Who is the Audience for EHR Evaluations?

When the "Footprints" task was initiated, the audience for the evaluations was not identified and specific evaluation questions had not been spelled out. It is clear from the papers that participants had very different notions with respect to who the audience is or should be. For some, the audience was the personnel of projects that the programs had funded; for others, it was the educational research community; for still others, it was primarily Federal decisionmakers, including executive and congressional watchdogs and funding agencies. Some participants assumed that the evaluations had a narrowly defined accountability purpose, documenting the extent to which progress had been made toward the attainment of the short-term goals that projects had been set up to achieve. Others assumed that the evaluation should be guided by a heuristic perspective and assess the extent to which NSF programs had funded projects that dealt with important issues, had contributed to the generation of new knowledge, and could be expected to improve educational practice over time.

Several conference participants emphasized the need for audience definition before adopting the evaluation questions and methodologies that seem most appropriate. This point was strongly emphasized by two discussants with considerable experience in conducting federally sponsored evaluations (Raizen, Chelimsky), and was also addressed by several other participants (Johnson, David Jenness, Yin, Boruch).

Audience definition is also a question that RED, and not the research community, must ultimately answer. What are the questions that the upcoming cycle of evaluations are supposed to answer, and whose questions are they:

- The program directors', to tell them how well all or some of the program goals have been met?

- The NSF policymakers', to help them assess the relative effects of programs now in place and perhaps identify new directions for program priorities?

- The educational research community, to alert them to the results, dissemination, and footprints of work funded in the past and perhaps needed directions for future grant applications and grant reviews?

- Or administrators in NSF and in oversight agencies, to tell them which programs had the best effects (payoffs)?

Furthermore, the audience may or may not be the same for every evaluation that is to be undertaken. Before a final evaluation design is selected, the audience question needs to be answered since it is unlikely that a comprehensive evaluation, which would meet the needs and interests of all potential audiences, can be designed within current budget constraints.

Can a Standard EHR Evaluation Model be Developed?

In his overview of the “Footprints” conference, Scriven stressed the desirability of using a consistent model across the board for all programs funded by EHR, because this provides a degree of comparability. Stake, on the other hand, argued in favor of using different models depending on the structure and goals of each program. Webb also pointed to the need for using multiple methods of inquiry in light of the large number of variables and complexities characteristic in educational research. Furthermore, while the suggestions that emerged from the “Footprints” conference tended primarily, but not exclusively, toward qualitative approaches, several suggestions, particularly Yin’s proposed analytic model, have a strong quantitative component. There are other ways in which quantitative approaches, such as sample surveys of project participants, e.g., teachers or administrators, could play a useful role.

The extent to which RED will decide to base its evaluation strategies for EHR programs chiefly on the suggestions of the “Footprints” conference participants depends of course on NSF’s ultimate decisions about the target audience and judgments about the types of information that this audience will require. For example, if costs and benefits are to be an element that should be considered in the evaluation, evaluation models quite different from those proposed by the conference participants, incorporating quantitative approaches that were not mentioned would need to be developed.

While there can be no question that a standard evaluation model would have great advantages, we do not visualize how it can be implemented, given the diversity of programs and the likelihood that different audiences might be targeted for various types of program evaluations. However, we have concluded from the examination of common conference threads that there may well be a set of core evaluation topics and questions that can and should be included in all evaluations. These are discussed in the next subsection.

Ideas and techniques that RED should implement for all evaluations include:

- Tracking selected program footprints or impacts;

- Archiving utilization information;

- Using portfolio assessment;

- Exploring the role of intermediaries; and

- Examining timing and extent of dissemination.

Tracking Selected Program Footprints. Most participants found the "Footprints" concept a useful one, although for many of them, "Footprints" is primarily a tool to be used for the construction of more elaborate evaluation strategies. But as a first step in the implementation of any of the strategies recommended at the "Footprints" conference, a comprehensive and coherent inventory of existing footprints is needed.

Several of the presenters came up with long lists of evaluation questions that an examination of footprints could answer and suggested possible sources for locating them. (The paper by Boruch, who focused on the Studies and Indicator programs, was most specific with respect to the latter.) As suggested by the participants and discussants, these lists need to be reviewed, so that for each program, a manageable, preliminary list of footprints and their sources for each of the four "effects" areas outlined by RED (effects on the profession, on other research, on educational practices, and on the funding agenda of other institutions) can be established.

While such lists will no doubt be modified as the evaluation task progresses, it is imperative to start with the compilation of a systematic, well-defined, and parsimonious set of footprints for each program that is to be evaluated and documentary and other sources where these footprints might be located.

Several of the conference papers provide a good starting point for these compilations, but a good deal of additional work is required. Particular attention should be given to sources and informants that commonly used bibliographic searches will not uncover (see Boruch's suggestions). It is also likely that relevant information can be located in program and project files, for example in applications for grant renewals, progress reports, or peer reviews. Once a first set of footprints has been compiled, it may be productive to seek reactions and suggestions for additional types and sources of footprints from selected policymakers and researchers who are active in a given program area.

The next step must be the bounding, classification, and ordering of footprints, along conceptually meaningful dimensions. Thus, the accumulation and classification of footprint data is a complex task, requiring both the casting of a wide net to capture "hidden" footprints, the setting of boundaries, and the creation of "Footprints" categories that will enable the evaluator to perform meaningful descriptions and interpretations of the data. Whether or not boundary setting should precede the data collection, or be done subsequently, is probably best decided on a program-by-program basis.

Depending on the audience and design, this initial data compilation will provide the basis for the following evaluation activities:

- A crude assessment of the program's visibility and potential impact in each of the four "effects" areas mentioned earlier;

- The selection of outcome indicators and other variables for the construction of a causal model based on partial comparisons (Yin);

- The decision to substitute a sample of projects for the universe in order to carry out analytic procedures with a more manageable data set (Raizen's proposed methodology for sampling based on a project typology seems especially useful); and

- The selection criteria for case studies if the evaluation design calls for this activity.

Because the choice of evaluation strategies may be dependent to some extent on the volume and characteristics of footprints that are identified, NSF may find it useful to undertake the compilations prior to finalizing evaluation designs.

Archiving Utilization Information. As was stressed by Boruch and pointed out by several other participants, there is at this time no mechanism in place to obtain systematic information about the use of data and research findings generated by EHR. Knowledge resides at the program and project level in professional publications (citations, other references, etc.) and in public policy documents (minutes of congressional hearings, speeches by officials, etc.). To sustain an ongoing evaluation effort based on footprints, the establishment of an archive where this information can be stored and accessed is of great importance. In particular, program and project staff should be required to provide periodic "utilization information" to this archive.

Portfolio Assessment. Another recurring idea dealt with the need to take a broader perspective and look at the entire educational research system and at funding sources other than NSF when evaluating program effects. Also, rather than looking only at areas where footprints might be found, several authors and discussants identified a series of evaluation questions that would provide a meaningful context for footprints, suggesting some kind of mapping or portfolio approach:

- Is the universe of projects funded by EHR a true reflection of the interests of the research community (David Jenness)?

- What would have happened if projects other than those for which awards have been made would have been funded (Johnson)?

- Why are there no footprints from a funded project and what can be learned by looking at unsuccessful or unfunded research (Webb)?

While some of these questions, according to Johnson, call for the evaluator to measure the immeasurable, it is evident that any evaluation of EHR programs would benefit from the more sophisticated approach of looking at EHR's "Footprints" programs in the broader context of the total science, mathematics, engineering, and technical education (SMET) research effort. This effort is funded by many sources besides NSF and carried out by researchers who have their own agendas, which influence how grant monies are expended and the extent to which performance bears a close relation to what was originally proposed in the funding applications (David Jenness, Boruch, Yin).

The questions raised by a number of participants addressed fundamental issues that the evaluation of the sizable and complex programs funded by EHR should consider:

- How well does each program target its awards?

- To what extent do programs address the right issues and respond to existing urgent needs for basic and applied research?

- Does the peer review process fund research stimulated by grantees' priorities for which they receive support from many sources?

- Do worthwhile proposals fail to obtain funding?

While NSF has instituted a mechanism for a broad review of these issues through periodic meetings of its Committee of Visitors and through the Expert Panels, a more systematic portfolio assessment is needed, based on an examination of funded awards, unfunded applications, funding activities carried out by other public and private agencies and an objective assessment of needs in the area for which the program bears responsibility.

One technique that might be useful in making portfolio assessments is a model proposed by Webb, represented on page 148, that uses a 2x2 matrix to address four key areas: what we have (or have not) learned from research supported by a program, the extent to which findings have been used, what problems have not been addressed by the program, and how the gap was filled. Webb limited himself to the RTL program when he developed this model and proposed specific types of studies for answering the questions raised. However, the model could be adapted for all or most EHR programs, since it goes to the core of issues that concern educational leaders as well as policymakers in funding and oversight agencies.

The Role of Intermediaries and Gatekeepers. Several of the papers have pointed to the important role played by intermediaries in acting as facilitators and gatekeepers in acquainting potential users (policymakers and practitioners) with research findings. Although this issue relates to some extent to dissemination, it should be examined in the "Footprints" context and needs to be considered for every EHR program that is being evaluated, although the types of intermediaries and the gatekeeping function they perform will differ widely.

In her paper, Christine Dwyer argued for a full-blown study of the paths and processes by which the Research in Teaching and Learning Program (RTL) influences educational practice, by examining the treatment of NSF-generated information by intermediaries and exploring the factors that determine transfer/nontransfer of this knowledge to practitioners (school personnel). The case studies that Dwyer proposes as a first step are exploratory in nature, focusing primarily on the intermediaries modus operandi, rather than systematic attention to the fate of EHR products. In her discussion, Raizen raised several caveats. In particular, she cautioned that intermediaries must be carefully selected and that not all intermediaries afford a valid test of information exchange. She also felt that rather than using the policies and practices of intermediaries as the starting point for case studies, it might be more useful to start out with some specific practice that looks as if it had been influenced by some assessed program and then trace back where the practice came from. Another approach that NSF may want to consider is to look at one major project within a given program to examine its treatment by relevant intermediaries (including some, such as museums,

Exhibit 1

Research Results	Applications	
Know	**Yes**	**No**
	What findings and information have been produced that have successfully solved a problem or fulfilled a need?	What findings and information have been produced that have not been applied to solve an important problem or fulfill a need?
Do Not Know	What critical problems or needs have not been resolved or refined by research findings and information?	What negative or poor applications have filled the gap in the absence of solid research findings and information?

whose main function is not service to education practioners), and examine the extent to which its findings did or did not reach the targeted audience. If carefully shaped so as to focus attention on the issue of concern to EHR, pilot studies of the role played by intermediaries could be very useful indeed.

Dissemination. There can be little argument that in many cases, the number of footprints is directly related to dissemination efforts on the part of investigators. NSF may want to investigate the extent to which the footprints that have been uncovered resulted from dissemination efforts by NSF program and project staff, and identify those dissemination techniques that have been most effective in yielding footprints. Initially, one or two case studies might be undertaken.

The many related issues, which the conference participants touched upon but did not develop, addressed the relationship between evaluation and dissemination. Several discussants (Raizen, Chelimsky, and Scriven) pointed out that dissemination is not appropriate for all research undertakings and is an expensive activity. Hezel, on the other hand, felt that evaluating the dissemination activities was a major task for the evaluation. There was also no thorough discussion about how to reconcile the need for early and widespread dissemination, which is emphasized in NSF proposal guidelines, with the time constraints imposed by evaluation and validation of project results, when projects are designed to affect educational practice and replication of successful projects is a program goal.

Scriven stated in his summation that although dissemination was included in the presentation and discussion of several conference participants, it was not a topic on the "Footprints" agenda and should be treated as an important but separate topic from evaluation.

Ideas and techniques that may differ with respect to various evaluations include:

- Need for causal attribution;
- Choice of evaluation methodology;
- Use of innovative analytic frameworks; and
- Use of innovative data collection.

Need for Causal Attribution. Those participants who tended to focus on the evaluation needs of Federal stakeholders (NSF, OMB, and Congress) and on the harder question of program worth felt that causal attribution had to be an essential ingredient of evaluations of federally funded programs (Scriven, Raizen, Chelimsky). In some cases impact attribution may also be important for program and directorate staff or the educational research community; in other cases, it may be more useful to devote resources to more extensive descriptive data for these audiences. The question of causal attribution was most fully addressed by Yin, who devoted his paper to the presentation of a new analytic technique to assess program effectiveness and make possible causal attribution of effects in the absence of controlled evaluation designs. Webb's paper also addresses the issue of attribution of effects. The recommendations of Yin and Webb are discussed in greater detail below (analytic frameworks).

Choice of Evaluation Methodology. In setting out the "Footprints" task, RED emphasized the need for finding new ways of evaluating the unique and innovative programs being supported in mathematics and science education and suggested that both new methodologies and new questions needed to be developed. While the participants presented many different ideas and differed on many issues, the one point on which there was agreement among the largest number of presenters and discussants was that the prevailing educational evaluation methodology, the service delivery model, is inadequate for the evaluation of many EHR programs and that viable alternatives do exist.

The alternatives offered took on many dimensions. At times nontraditional was equated with qualitative, and, therefore, traditional was associated with quantitative methods. Some participants (Barley and Mark Jenness) defined nontraditional methods as those that emphasize the interests of project clients and other local stakeholders and use negotiation as the major evaluation tool. While Stake questioned the use of any systematic evaluation method (because of the dominance of the political and administrative context in which the programs operate), most participants offered nontraditional evaluation strategies using both improved new approaches to educational evaluation and traditional scientific methods from other fields, especially ethnographic and cultural studies.

Indeed, the description of proposed nontraditional approaches led one discussant (Phelps) to comment that "they all model what should be and is good evaluation practice. They are only nontraditional in the sense that in the Federal Government they are not often carried out."

In his comments, Scriven took exception to the widely expressed need for new methodologies. In his words, he found himself in the unfamiliar position of defending the status quo. He felt that the arguments for needing a new approach were mostly aimed at what is really a straw man and faulted the NSF's procurement policies, rather than shortcomings of the methodology. He asserted that the agency had not tapped into the best available evaluation practices, which are a kind of eclectic amalgam of qualitative and quantitative methods, carried out by experienced and sophisticated evaluators.

Taken together, the comments by conference participants suggest that while RED should continue to encourage the development of innovative methodologies, there is no need to rely solely on methodologies developed from scratch. While it may be necessary to do so for the evaluation of some programs, for others (for example the RTL program) the "eclectic mix" recommended by Scriven may be most appropriate. Furthermore, there presently exists a number of fully or partially developed models that are not based on the service delivery approach. A first step should be to explore the alternatives with the goal of adopting (or adapting) some of the quantitative and qualitative approaches that already are used in our own and other fields. The ideas and techniques proposed by the "Footprints" authors may be considered nontraditional with regard to common practice in federally funded evaluations, but many of them are based on data collection and analytic approaches with established histories and credibility.

Alternative Analytic Frameworks. Three of the conference papers (Yin, Webb, Barley and Mark Jenness) focused on innovative techniques for developing analytic frameworks for EHR evaluations.

Yin's objective was to use footprints to establish a causal link between program activities and observed outcomes through the use of a rigorous technique that would be an acceptable substitute for experiments or quasi-experiments used in traditional service delivery-based models, which are inappropriate for most EHR programs. The usual characteristics of grant programs are that the intervention carried out by grant-funded projects is weak or small, relative to the impact of interest; the intervention is not part of a formal research design; and extensive time or resources are not available for the research effort. Given these problems, experimental designs must be ruled out. Database analyses are primarily descriptive and do not permit causal inferences.

Instead, Yin recommends a new methodological strategy, which aims at making "multiple, partial comparisons" instead of imposing a singular research design in carrying out an evaluation. Unlike traditional evaluation designs, this method can be used when evaluators have no control over the intervention or when the interventions do not meet the statistical requirements of any of the "traditional" designs. Partial comparisons can enable investigators to offer causal inferences by using single components (specific project effects) as the main unit of analysis. The larger the number of positive inferences that can be supported through these partial comparisons, the stronger the argument that positive results were produced and the stronger the conclusion that the program under evaluation produced them. This strategy requires the evaluator to identify and collect data, in effect footprints, that can satisfy as many partial comparisons as possible. Outcome data from projects funded by the program are the relevant input for each partial comparison, and the instruments needed to collect these data will vary. The AAT program was one for which he felt this approach would be especially suitable.

The paper presented by Webb presented several strategies for the analysis of footprints. Especially useful was his suggestion about dealing with the very large number of footprints that some programs are likely to yield (he focused on the RTL program that to date has funded more than 200 projects). One of the issues often raised by critics of qualitative approaches is that investigators are very good at collecting a great deal of interesting data but have not developed rigorous methodologies for their interpretation. Webb proposed a generalizability analysis to substitute the study of a sample of projects, selected at random, that would yield a cross-section of projects and provide a good description of the program as a whole. In her discussion, Raizen proposed an alternative to random sampling of projects, recommending instead a two-stage approach, with some initial grouping of projects along common dimensions, such as problem addressed, or approach taken, and subsequent sampling within each of these groups. Raizen emphasized that the groupings would have to be thought through very carefully, but if this was done, the sample used for analysis would be greatly superior to one obtained through random sampling.

Both Webb and Yin sought to build comprehensive evaluation models to shed light on the value of programs, address the issue of utilization of findings, and answer questions of causality. Webb's approach, discussed earlier, used a 2x2 matrix to examine the extent to which research has yielded findings that were used to solve educational problems. Yin's model incorporated the concept of rival hypotheses to test the causal link between research findings and the adoption of educational innovation. His proposed analytic technique, partial comparisons, appears promising. Considerable work on partial comparisons has already been done by Yin for other agencies.

The framework proposed by Barley and Mark Jenness is based on a different premise. They believe that the main goal of evaluation is formative and aimed at project and program improvement. Their proposed cluster evaluation concept and techniques for its implementation have been tested, with support from the W.K. Kellogg Foundation for formative but not for summative assessments. Barley and Mark Jenness recommend its use for summative program evaluation through the creation of samples of retrospective clusters, consisting of completed projects, based on regional or topical sampling frames. A "cluster evaluator" would work with directors and other project staff to negotiate a set of common cluster outcomes and collect both qualitative and quantitative data from a variety of sources using various techniques. Some common cluster instruments, used across projects to collect consistent data, can be created for the data collection. Scriven has forcefully argued against this approach, pointing to the credibility and objectivity issues that its use would create for a summative evaluation. A more limited use of this technique, confined to data collection only and discussed later in this paper, might be considered.

Incorporation of all or part of Webb's and Yin's models and techniques in an evaluation design would greatly increase the sophistication of footprints analyses. Both models would require substantial data collection, in particular a fairly complete mapping of all efforts sponsored by public and private agencies that are directed at the strengthening of mathematics and science education and recruitment. This mapping would be a difficult and time-consuming undertaking; again, a sampling approach seems indicated. After data have been collected, the suggested models for attributing specific outcomes to EHR programs can be fleshed out.

Yin sees the need for further methodological development before the partial comparisons technique can be tried for the evaluation of NSF programs. Key outcome measures (for example, new ideas for research or practice) have to be developed. To pinpoint effects traceable to NSF-funded programs, case studies need to be conducted of funded investigators and the projects they undertake, so as to develop information about how grantees merge various sources of support to carry out their research projects. The list of partial comparisons needs to be expanded to be suitable for EHR programs, and pilot testing should be done to assess the efforts and costs required. But if EHR sees the need for in-depth assessments of program outcomes, these methods are certainly worth exploring further.

Innovative data collection. Several of the papers, especially those by Boruch, Johnson, and Barley and Mark Jenness, contain innovative suggestions for data sources and data collection techniques that could be explored. Boruch, who focused his discussion on RED's Studies and Indicators programs, offered an extensive list of possible sources of references and uses going beyond the commonly used citation counts and publications in refereed journals by high-quality publishers. He suggests professional recognition through awards and prizes, presentations in professional and public forums, and popular press or media coverage. He also recommends scanning press and agency reports that have used a study without directly acknowledging the source, direct observation of public meetings where studies are discussed, and self-reports by project staff, usually the principal investigator. Peer reviews, review panels, and the knowledge of seasoned staff in foundation grant programs and Federal agencies are other good sources. Boruch further pointed to somewhat more remote effectiveness indicators, such as contributions to research methodology and data production methods. He recognizes that the systematic accumulation of this information may well be a monumental task, best carried out in an academic setting where graduate students constitute an affordable labor source.

Clusters could be a practical data and information collection resource, standardize evaluation questions. The RTL program is a good candidate for this approach. Using a common data collection instrument for projects in a given cluster would standardize evaluation questions and facilitate the collection of a common core of data for a given program. This approach might be useful for the RTL program.

Recommendations

The reason for initiating the "Footprints" task was to develop some nontraditional approaches to evaluating ERR programs, which, because of their organization, goals and support structure, are not easily amenable to being examined using the typical Federal evaluation model. The varied experts whose ideas were tapped as authors or discussants have provided NSF with a long list of ideas from which to choose in approaching these evaluations. In this paper, we have selected for more extensive discussion those suggestions that we felt were especially promising. While several useful methodologies and frameworks for assessing programs' worth have been offered, we believe that the most useful contribution that the conference (and this paper) may have made is the identification of the common core of activities that we have outlined: tracking selected program footprints, portfolio assessment, the role of intermediaries, and the relationship between evaluation and dissemination. We also feel that the identification of evaluation audiences is of paramount importance before specific evaluations are designed.

What happens next depends on a number of steps that EHR itself must take; steps that involve possibly investing in the fuller development of some of the alternatives offered, as well as setting priorities among audiences and questions to be addressed. Given the innovative nature of some of the proposed procedures, small-scale pilot testing would also be advisable. We have identified several techniques that EHR may want to consider in planning upcoming evaluations for specific programs, and some methodological tasks that might be undertaken prior to the adoption of final evaluation designs. These include:

Develop a System for the Collection of Footprints from NSF Program and Project Files. Several discussants pointed to the role that NSF itself, as well as funded projects, must play in accumulating footprints. These recommendations have been discussed earlier. Written requests for copies of reports and other types of information, telephone inquiries about findings, invitations extended to program and project staff to participate in activities where program-generated information is to be discussed are not systematically documented at the program level. Boruch saw the need for an NSF program archive; other presenters emphasized the role of the project director. At present, available information is largely anecdotal and decentralized. As part of the current EHR effort for database creation, it may be possible to generate systematic Footprint data at the program and project level.

Develop a Methodology for Portfolio Assessment. The Webb matrix represents one possible approach; Yin's "rival hypothesis" also addresses the issue. But EHR needs a comprehensive strategy to carry out this assessment for all its programs.

Conceptualize and Pilot-test the Intermediary Function as it May Apply to all EHR Programs. Once appropriate intermediaries have been identified for several programs, it may be useful to adopt Raizen's strategy and examine in a pilot test the role played by these intermediaries with respect to one or more products that resulted from these programs.

Clarify EHR's Policy with Respect to the Connection between Evaluation and Dissemination. Here, too, it would probably be useful to look at some actual dissemination practices and examine their effectiveness as well as their relation to evaluation efforts and outcomes.

If the Causal Attribution of Program Effects is to be Included in the Evaluation, Develop and Pilot-test the Partial Comparison Methodology for the Program to be Evaluated. As suggested in the earlier discussion, it is not obvious that the model and analyses proposed by Yin will be appropriate for all EHR evaluations. When they are used, considerable methodological development and pilot testing will be needed, as Yin himself has emphasized.