New Methods For Evaluating Programs In NSF'S Division Of Research, Evaluation And Dissemination

Robert K. Yin
COSMOS Corporation

Basic Nature of Grant Programs and Purpose of This Paper

The National Science Foundation (NSF) sponsors many programs in science, engineering, and mathematics (SEM) education. All of these programs are "extramural," in that NSF makes awards to some performing organization—generally a university or nonprofit organization. The award is usually a grant award, administered under conditions specified in Grants for Research and Education in Science and Education (NSF, 90-77, October 1992). (The programs also make contract awards and enter into cooperative agreements, but these are a very low proportion of all awards and are not the subject of this paper.)

With a grant award, the performing or grantee organization is supposed to conduct a "project." These funded projects become the collective entity known then as the "program," and individual NSF programs routinely issue reports on the nature of these funded projects. (In many circumstances, the work done under these funded projects may not be readily delineated from work supported by other funded projects simultaneously received by the grantee, but this topic also is beyond the scope of this paper.)

The challenge addressed by the present paper is to develop better methodologies for evaluating programs consisting of this sort of infrastructure. Three NSF programs in particular were used as background information for this challenge:

- **Applications of Advanced Technologies Program** ("AAT" program);
- **Policy-Related Research: Studies Program** ("Studies" program); and
- **Policy-Related Research: Education Indicators** ("Indicators" program).

The paper only aims at developing preliminary ideas in this methodological direction and is not intended to be a complete prescription or even operational set of guidelines for carrying out an evaluation. Rather, the goal is to describe why such new methodologies are needed, and then to point to the further methodological work to be done that will lead to the creation of these better methodologies.

Potential Conflicts Between Grant Programs and "Standard" Program Evaluation Methods

The Standard Program Evaluation Model

The need for new methods derives from the potential inappropriateness of the standard program evaluation model as it might be applied to a grant program. Exhibit 1 contains a simplified version of the standard evaluation model. The model puts heavy emphasis on the iden-

"The need for new methods derives from the potential inappropriateness of the standard program evaluation model as it might be applied to a grant program."

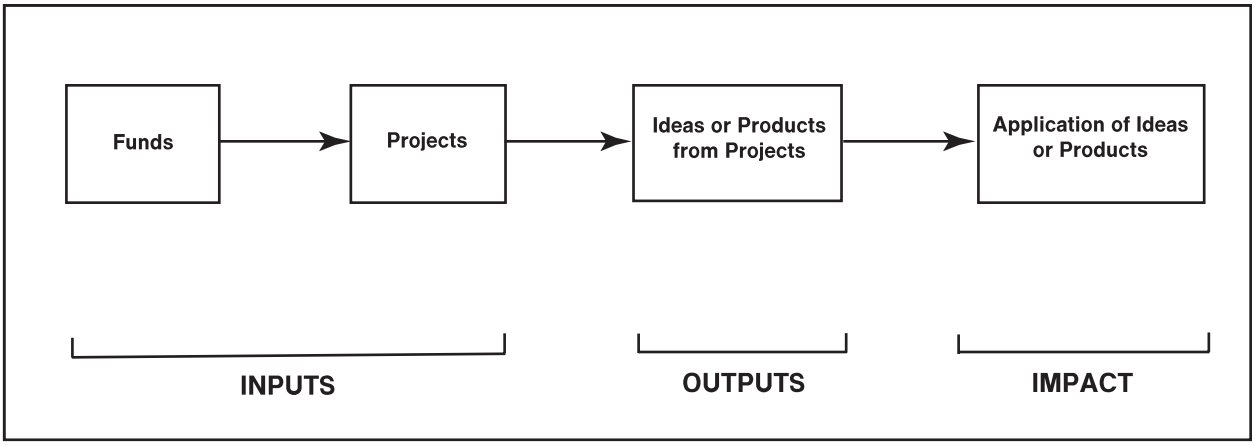

Exhibit 1 "Standard" Program Evaluation Model

tification of real-world impacts. In the SEM education field, such impacts would be expected to occur in actual school systems (K-12 or universities), involving actual teachers and classrooms, and therefore involving actual students. Traditionally, the model also puts heavy emphasis on defining the impacts in quantitative terms. Ideally, the model would like to help policy makers understand how many classrooms and students or teachers were impacted, and to what quantifiable degree, by investing in a particular NSF program.

Attempts to implement the model usually begin with data being collected about the individual projects. The projects may have led directly to applications in the field—and hence may have produced impacts that can be measured. However, if the projects only produce new ideas that are not carried into the field, the model may not be useful. Similarly, the user of standard evaluation data collection methods will encounter difficulties if the impact in the field: a) occurs over a long period of time (say, 10 years) after the ideas were first produced by the project—a commonplace time lag in SEM education; or b) is difficult to attribute because of the relatively small size of the NSF program investment—also a commonplace occurrence because NSF's investment may be in the millions of dollars, whereas the education system of the United States operates at the level of tens of billions of dollars. In either situation, the resulting impacts may be considered overdetermined, and attributing them to NSF-funded projects is hazardous at best.

As a general rule, because education is largely a state or local matter (grades K-12) or a university matter (postsecondary), Federal initiatives must be relegated to extremely minor roles. For instance, the Studies program lists as its major goal the strengthening of SEM education in the United States. Such an impact is very hard to trace, however, given that the program operates with an annual budget of less than $5 million. Similarly, of the three programs, the largest is the AAT program, which supports $10-20 million of funded projects annually in an educational technology market worth at least hundreds of millions (if not billions) of dollars.

At a programmatic level, the interpretation of the results of a standard evaluation also may be little more than the aggregate of all of the project-level results. Strategic considerations pursued by programs—e.g., to overinvest in certain areas of high priority, or to make a few high-risk awards, or to follow any portfolio criteria—tend not to be covered well by the standard evaluation model, as traditionally practiced.

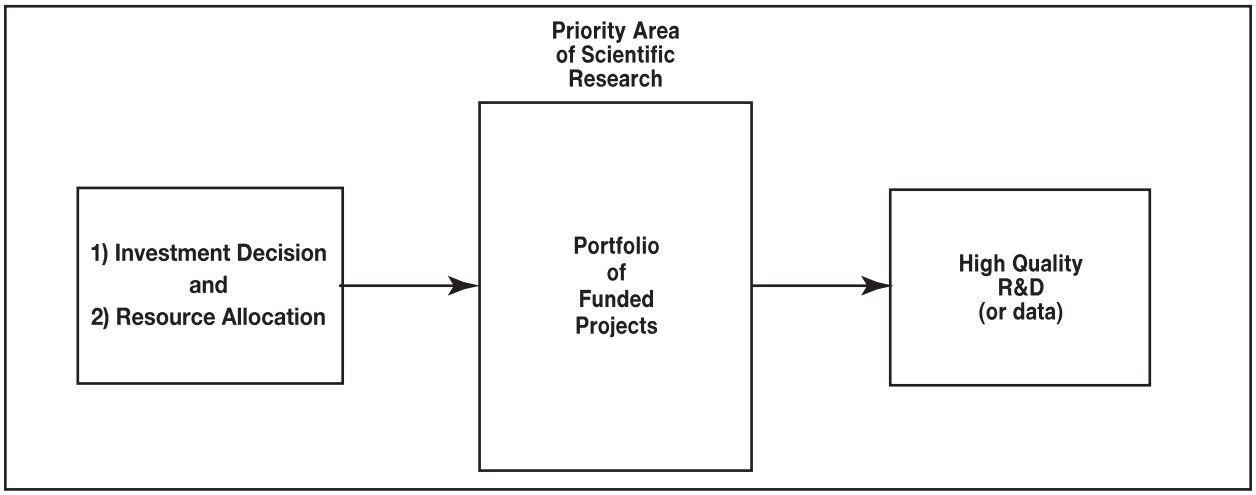

Exhibit 2 A Grant Program Model

A Grant Program Model

The standard evaluation model does not reflect well the way that Federal grant programs are created, or how the staff or sponsors of grant programs usually strategize about their programs. Exhibit 2 contains a simplified version of how the program might be conceptualized by its staff or sponsors, using a grant program model. Essentially, a public commitment has been made to support R&D in a pre-identified priority area of scientific research. The role of NSF, as a sponsoring agency, is to make these awards in as rigorous and utilitarian a fashion as possible. The main subsequent concern is whether the research was completed in a high-quality manner.

In most grant programs, the grants are used to support basic research. But even where applied research is the main subject of a program, this same type of thinking has traditionally been followed for two main reasons. First, the mandating legislation may contain no specific goals (for example, none of the three NSF programs have specific legislative mandates), and none may have been articulated beyond the statement of need for investing in the area. Second, the award characteristics of a grant mitigate against other ways of thinking. Grant awards deliberately permit grantees to make reasonable adjustments in a project as it starts up and is implemented. Indeed, the purpose of a grant is not to limit an investigation to a rigid design, but to encourage the investigator to make the best choices leading to high-quality R&D. Further, the grant award is considered important in attracting proposals from highly capable investors, who have

traditionally been able to take advantage of the independence of grant award conditions to create inventive results.

In the grant program model, the notion of quality would include such criteria as: 1) advancing the state of understanding about a topic ("making a contribution"), 2) developing a framework or foundation for further research on a topic, and 3) far exceeding the standards of a field or academic discipline. Quality may not necessarily include such criteria as relevance to immediate problems, much less having an impact on them.

"... some scholars reduce the likelihood of failure by performing new studies that are one step ahead of their awards."

When a grantee fails to produce high-quality R&D, the major consequence is that—in the long run—the grantee will find it increasingly difficult to obtain new grants. However, other than strictures regarding fraud, waste, and abuse, it is not incumbent for the grantee to "perform" productively on any given grant award. On the contrary, the underlying philosophy is that much new research will fail, and that the nature of research involves a high incidence of failure. In fact, the grant mechanism was designed in part to accommodate this aspect of the scientific enterprise.

Competitive scholars, of course, will always find a way to produce a gain from every funded project. A minor publication, a new descriptive understanding, or a methodological lesson may have to compensate for the failure to complete the original project as proposed. As another variation, some scholars reduce the likelihood of failure by performing new studies that are one step ahead of their awards. Their new proposals therefore contain proposed inquiries whose outcomes are already known, though not yet published or shared with colleagues—and therefore increase the probability of getting a grant award.

The grant model, however, clashes with the traditional evaluation model. The grant model gives little attention to impact. At the same time, high value is placed on quality—which in turn is generally ignored by the quantitative orientation of the traditional evaluation model. In addition, unlike the traditional evaluation model, the grant model highlights the portfolio of projects and incorporates strategic investment goals that are not just the aggregate of all individual projects. For instance, the AAT program prides itself in being a "high-risk, high-gain" effort. In other words, the hope of the program administrators is that a few of their projects will produce scientific breakthroughs, even though the majority of the projects may not lead to significant advances in knowledge. The grant model accommodates this strategic objective more readily than the traditional evaluation model.

Why Evaluation Is Needed

Public investments in grant programs, whether in support of basic research, applied research, or R&D more generally, necessitate the assessment of external benchmarks of progress. Most commonly, the evaluation of a grant program is put into the hands of an expert panel, which may be organized as a "visiting committee" or operate under some prestigious sponsorship such as the National Academy of Sciences. NSF-SEM education programs have been subjected to these types of evaluations as well as numerous other administrative reviews. The challenge is not to displace these efforts, but to ascertain whether formal evaluation methods can complement them.

A New Evaluation Strategy

Formal evaluations can, in fact, be complementary, if only the methods used to conduct them are modified. The modifications are needed to make evaluation applicable to situations such as these occurring in the R&D grant program, in which:

- The intervention is weak or small, relative to the measurable impact of interest;
- The intervention is not a part of a formal research design, because the intervention was not designed to suit the needs of evaluation, but rather to suit policy-related or real-life needs; and
- Extensive time (five years or more) or resources (millions of dollars) are not available to support the needed evaluation effort.

To deal with these conditions, COSMOS's ongoing research has been developing a new methodological strategy (Yin, 1993; and Yin and Sivilli, 1993). The main feature of this new strategy is that it aims to make multiple, partial comparisons instead of imposing a singular research design in carrying out an evaluation. The new strategy offers the opportunity to collect diverse data and to target multiple inquiries in lieu of an overarching research design. The new strategy and how it modifies the traditional evaluation model appears directly related to the evaluation of R&D grant programs.

Exhibit 3 summarizes the traditional evaluation model and its varieties, also showing the niche filled by the proposed new strategy. Randomized clinical trials ("true" experiments), quasi-experiments, and database analyses have all been used in the past as traditional evaluations. The U.S. General Accounting Office (1992) has developed a meta-analytic approach of synthesizing data from these different strategies. The proposed new strategy presents an alternative—filling the gaps between these strategies.

The exhibit shows that when research investigators have no control over the intervention, and when the interventions are not even designed to suit a research design, the need is for some new strategy more powerful than mere database analyses. The new strategy will make some causal inferences possible, even though these will not be nearly as potent as those in quasi-experiments or clinical trials. However, the new strategy may be more generalizable and less costly than quasi-experiments or clinical trials. The new strategy has six features:

- The use of partial comparisons, based on multiple "partial" designs;
- Designation of each single component of a comprehensive program—rather than the program as a whole—as the main unit of analysis (therefore leading to multiple sets of partial comparisons, if a program had several components);
- Greater emphasis on the use of proximal rather than distal outcomes where interventions are of low strength or "dosage;"
- Explicit assessment of the "process" logic of an intervention;
- Replication across multiple components or programs where objectives are similar; and
- Triangulation about key events by using multiple measures.

"... the new strategy may be more generalizable and less costly than quasi-experiments or clinical trials."

Exhibit 3 Evaluation "Niches"

Evaluation Strategy	Characteristics of Evaluation Situation				
	Researcher Control over Intervention	Interventions Intended to Suit a Research Design	Casual Interpretability	Generalizability	Cost
RANDOMIZED CLINICAL TRIALS	Yes	Yes	True comparisons and causal inferences possible	Trials limited to narrow client populations	High in cost
QUASI-EXPERIMENTS	No	Yes (except random assignment)	Causal inferences possible	Trials limited to narrow client populations	Moderate in cost
PARTIAL COMPARISONS	No	No	Series of partial comparisons possible, ruling out key rivals	Compaisons cover moderate range of client populations	Moderate in cost
DATABASE ANALYSES	No	No	Poor comparisons and causal inferences possible	Databases cover full range of client populations	Low in cost (data already collected)

Of these six features, the most innovative and important deals with partial comparisons, and the remainder of this paper therefore suggests how this feature might work in evaluating a program like the illustrative three programs of NSF.

Application of the New Strategy

Exhibit 4 lists an illustrative set of partial comparisons. The comparisons are considered partial because none alone provides definitive causal evidence about the outcomes of a program. However, each partial comparison is intended to support a positive inference about the program and its outcomes. Thus, the more partial comparisons that an evaluation can cover (and these partial comparisons go beyond the 18 listed in Exhibit 4), the more compelling the argument can be made that: a) positive results were produced, and b) the program under evaluation produced them. The goal of the new evaluation strategy is therefore to identify and collect data that can satis-

Outcomes-Only Comparisons

1. The program performed better than at earlier time (pre-post).
2. The program performed better than another program (cross-section).
3. The program performed better than broader group of programs (cross-section).
4. The program's performance trend is in desired direction (time series).
5. Outcomes appear faster or better than expected.
6. Outcomes exceed initial goals or objectives.
7. Outcomes exceed established standards.

Process-Only Comparisons

8. The program implemented a new set of activities, not previously conducted.
9. The program improved an existing set of activities.
10. The program staff can describe how the program differs from previous policy or practice.

Causal Interpretation

11 The program staff can provide a compelling explanation for a documentable chain of events.
12 Ditto external observers
13. Ditto a key informant (insider)
14. The pattern of outcomes is uniquely related to the program.
15. The intervention is uniquely related to some infrastructure, in turn related to the outcomes.

Rival Interpretations

11R. The program staff can provide rationale for rejecting explanations:
- general climate
- competing programs.

12R. Ditto external observers
13R. Ditto a key informant
14R. Ditto pattern of outcomes
15R. Ditto infrastructure

Policy Analyses

16. Magnitude of positive outcomes far outweighs costs of program.
17. Outcomes achieved for the first time in this program.
18. Outcomes generate support for further desirable action.

Exhibit 4. Illustrative Partial Comparisons

fy as many of these partial comparisons as possible. The strategy provides flexibility because the relevant data for each partial comparison and the instruments needed to collect those data may vary. Further, no singular research design is being relied upon; rather, the final evaluation will consist of multiple, partial designs.

A critical subset of the partial comparisons is the explicit consideration of rival interpretations. Unlike database analyses, the new strategy encourages and accommodates the collection of evidence to test such rivals. The identification and selection of rivals is not easy (McGrath, 1982). However, the more the rivals are shown to be untenable, the greater the credibility that can be given to the target program's effects. To this extent, the new strategy should produce more definitive evidence than database analyses.

For the R&D grant program, the application of this new strategy yields a modified model of the R&D program, shown in Exhibit 5. This model shows that an evaluation can go beyond the grant program model (Exhibit 2) and assess the production of new ideas as a legitimate program outcome. These new ideas would be considered legitimate payoffs from any of the three NSF programs. For instance, the AAT program aims at producing new ideas demonstrating proof of concept, the Studies program aims at policy-relevant ideas, and the Indicators program aims at benchmarks reflecting educational progress. However, the model also falls short of the traditional evaluation model (Exhibit 1) in that it does not attempt to deal with program impacts.

Exhibit 6 shows how the modified model can be augmented to incorporate rival interpretations. Two such rivals are shown, although others might also be relevant. The Rival 1 hypothesis suggests that other funded projects produced the same valued ideas; the Rival 2 hypothesis suggests that other programs would have supported the same funded projects in the absence of the targeted program.

Immediate Needs for Developing the New Strategy

This new evaluation strategy cannot be put into place at the current time. Further evaluation or methodological research is needed to refine the strategy and make it operational. As a result, this paper concludes with recommended methodological steps, and not an actual plan for evaluating a real-life program.

The first recommendation is for the development of "measures" of the key program outcomes—new ideas (for research or for practice), influence on policy decision making, and capacity-building of the performer community (where relevant). Conceptually, any measures of new ideas should represent new concepts and new ways of thinking about a problem or situation. Similarly, influence on policy decision making should represent the incorporation of ideas into new decisions. Finally, capacity-building should represent improved skill levels and performance by appropriately trained personnel. Operationally, new ideas, impact on decision making, and capacity-building have generally been identified through peer review panels, such as committees organized by the National Academy of Sciences. Determining whether alternative measures can be developed is the objective of this first recommendation. For new ideas for applications, for instance, the AAT program's operationalization of "proof of concept" is already a

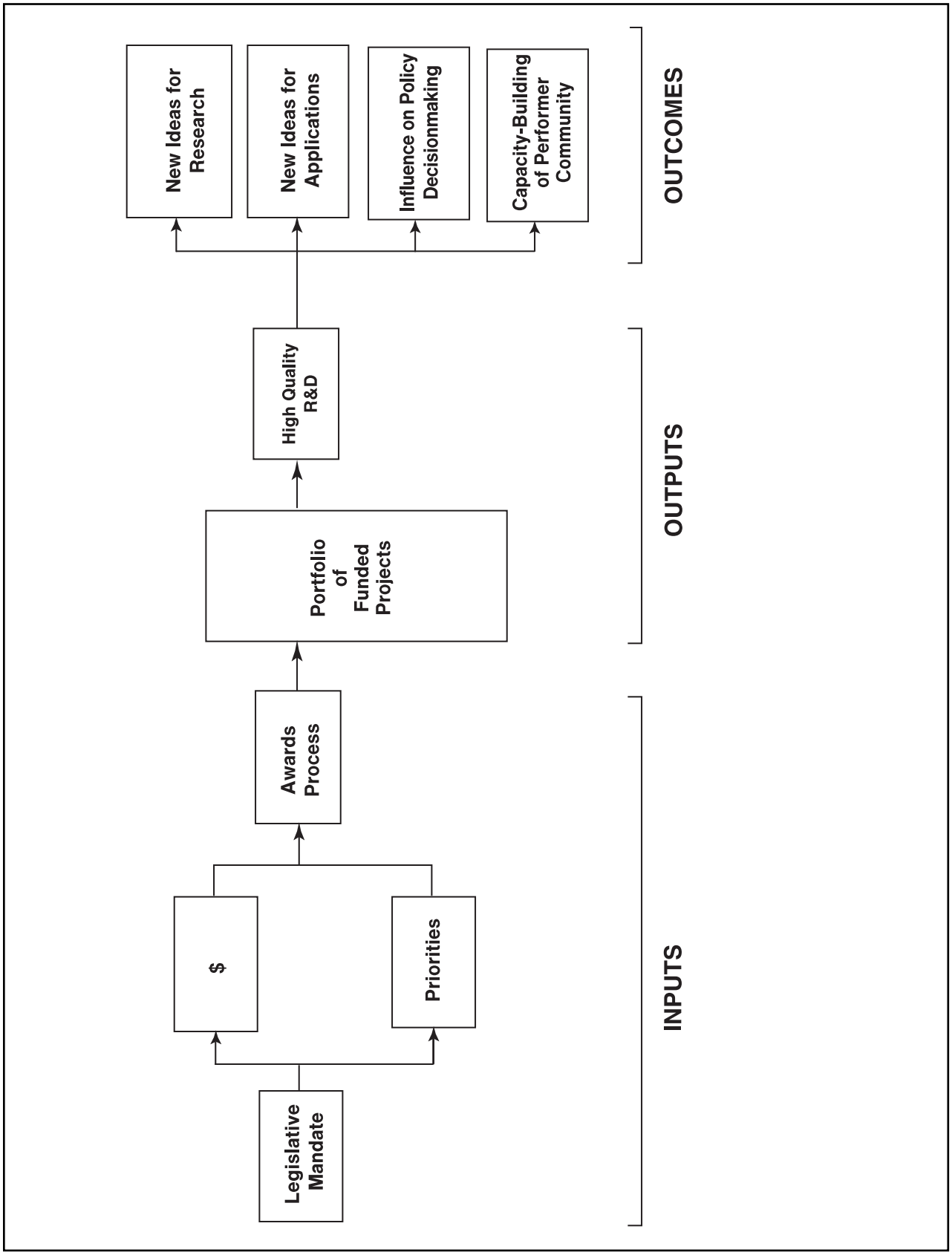

Exhibit 5. A More Practical Evaluation Paradigm Applied to an R&D Program (I)

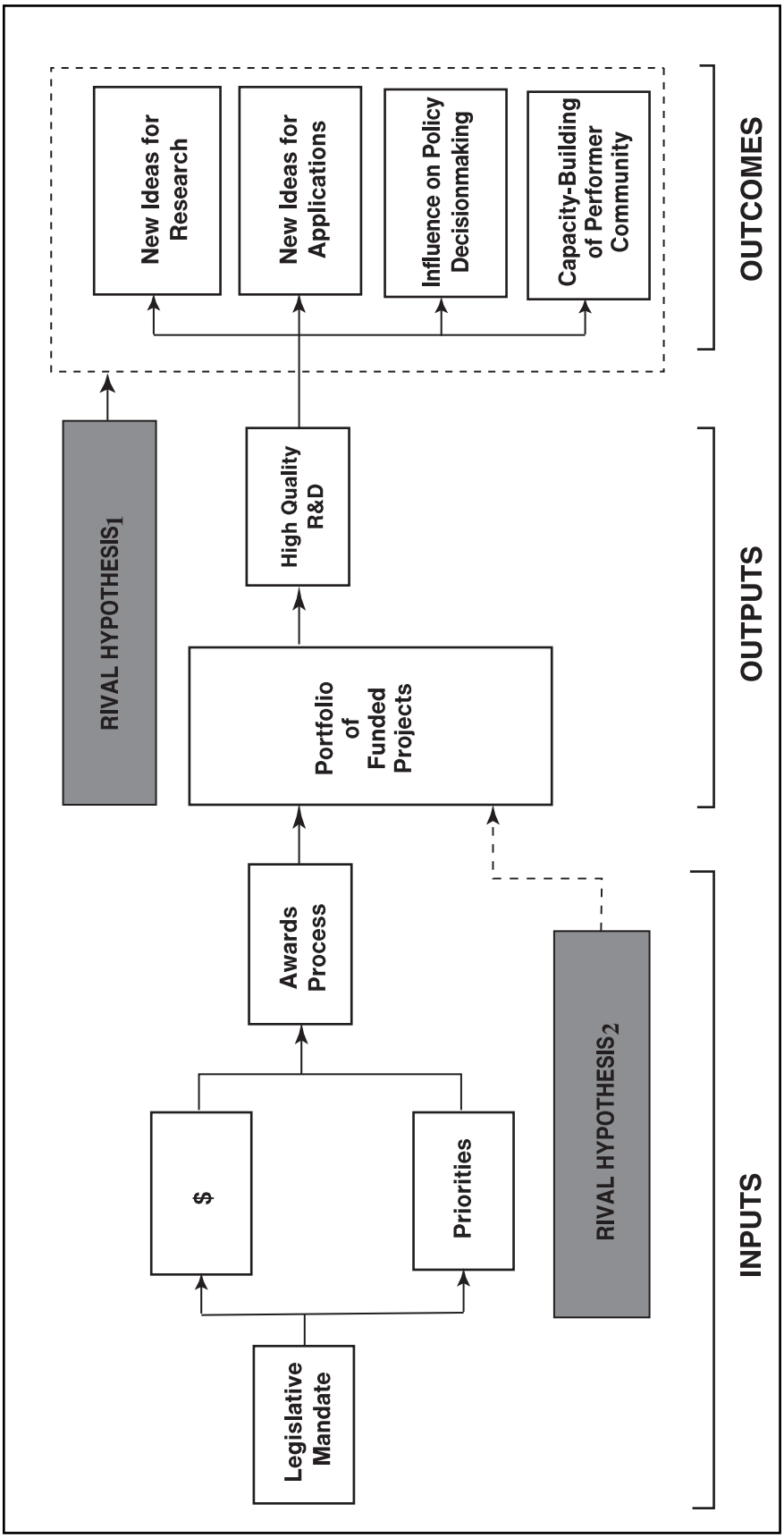

Exhibit 6. A More Practical Evaluation Paradigm Applied to an R&D Program (II)

promising approach that should be explored further as a methodological advance.

The second recommendation is to develop designs for conducting case studies of funded investigators and the projects they undertake. These investigators may be able to report or demonstrate how they have blended different sources of funds to make different projects or different findings possible. Such patterns might provide clues about the importance of the targeted program, compared to other sources of funds—thereby helping to unravel Rival 2 in the previous example (Exhibit 6).

The third recommendation is to extend the logical list of partial comparisons in Exhibit 4. A comprehensive list is needed, even if any given evaluation can only cover a subset of the list.

Finally, some testing needs to be done to assess the level of effort and costs of undertaking partial comparisons. Exhibit 3 assumed that these costs would be moderate, compared to the costs of conducting randomized clinical trials. However, actual data about the costs would be extremely informative. Some comfort may be derived from an earlier effort (Fitzsimmons, et al., 1992) that managed to track causal program relations within a reasonable time frame and cost limit. This earlier effort did not follow the proposed methodology but did cover a roughly similar scope, evaluating NSF's Coordinated Experimental Research in Computer Sciences (CER) Program.

Summary

The evaluation of ongoing Federal programs—in mathematics and science education and related research—is a challenging problem. The programs already exist, have been operating for some period of time, and were not designed to be part of formal evaluations. An evaluator must therefore address these programs without assuming the ability to manipulate key experimental or treatment conditions.

Traditional evaluation designs do not serve well under these circumstances. As a result, new evaluation strategies are needed. The present paper deals with this challenge by proposing a new strategy of partial comparisons. This new strategy entertains and deliberately seeks to investigate rival explanations and threats to validity. However, the strategy does not assume the creation of a singular evaluation design to deal with all rivals (as do traditional designs). Rather, the total evaluation of a single program will consist of multiple substudies—each potentially using different designs and sources of evidence as relevant.

This paper demonstrates, in a preliminary manner, how the new strategy would be relevant to typical NSF programs in mathematics and science education such as the Applications of Advanced Technologies Program, the Studies Program (policy-related research), and the Education Indicators Program (policy-related research). The paper concludes by identifying the needed methodological work before the strategy can be considered a truly competitive alternative.

"... some testing needs to be done to assess the level of effort and costs of undertaking partial comparisons."

References

Fitzsimmons, SJ., et al. 1992. *An evaluation of NSF's Coordinated Experimental Research in Computer Sciences (CER) Program.* Cambridge, Mass., and Washington, D.C: Abt Associates Inc. and COSMOS Corporation.

McGrath, J.E. 1982. Dilemmatics: The study of research choices and dilemmas. In *Judgment calls in research*, eds. J.E. McGrath, J. Martin, and R.A. Kulka. Beverly Hills, Calif: Sage Publications 69-102.

U.S. General Accounting Office. 1992. *Cross design synthesis: A new strategy for medical effectiveness research.* Washington, D.C: GAO/PEMD-92-18.

Yin, R.K. 1993. Evaluation design: Breaking new ground. Unpublished paper. Washington, D.C: COSMOS Corporation.

Yin, R.K., and Sivilli, J.S. 1993. Evaluation of gang interventions. Paper presented at National Institute of Justice's Fourth Annual Conference on Evaluating Crime and Drug Control Initiatives, June 28-30, Washington, D.C.

Thank you for the opportunity to react to the papers. Coming from an evaluation office charged with producing evaluation reports to inform policy and legislation for the elementary and secondary programs in the Department of Education, I appreciate the clear thinking that has gone into the writing of these papers. The presentation today lifts our sights beyond looking at our day-to-day evaluations in the traditional way.

Our problem in program evaluation studies, and I'm sure this is shared with the National Science Foundation, is that our evaluations are very much tied to the legislative cycles, to budgetary needs, and to looking at administrative changes that have to go on in programs. If they don't do that, they usually don't make it beyond the prospective stage. We rarely have use for studies for which we can't see immediate payoffs.

Further, we must work within some important limitations. Our funding is often dependent on a particular program or a congressional mandate to investigate a particular program. Chapter 1 presents a good example. Because we have a line item for evaluation in the Chapter 1 compensatory education program, it's little wonder that most of the activities in my office concern Chapter 1 and look at issues involving disadvantaged students. At the same time, we need to avoid getting stuck in a rut, relying on boilerplate methodologies when some radical rethinking is really needed. However, currently there is no demonstration authority in the largest of the Department of Education's elementary/secondary programs, Chapter 1. This means that our work is dependent on finding naturally occurring examples of effective practices and programs. Yet we realize that the field desperately needs new approaches to replace the low-level basic skill and drill models that currently prevail. These constraints lead us to take opportunities where we can find them.

Let me share some examples of using opportunities. When sufficient funds were unavailable to launch a full-scale national study looking at math and science programs for gifted and talented students, we scaled back to case studies. These case studies were done by Cosmos, Robert Yin's company. To limit the field of possible sites—we could have gone to hundreds and hundreds—we decided to focus on projects that served disadvantaged students. This resulted in a study that has contributed in several ways to refocusing the Federal effort on assisting the disadvantaged. The study findings were used to craft priorities and selection criteria for both Native American education and the Javits Gifted and Talented program. The study encouraged other work, spurring us to look at strategies from gifted and talented instruction that could be applied to the regular classroom and to examine the impact of these alternatives to conventional wisdom regarding educating disadvantaged students.

We try to stretch our resources and broaden the scope of our evaluations to examine the larger context for Federal programs, rather than always looking program by program. For example, we are currently competing an evaluation contract to examine the Eisenhower Regional Math and Science Consortia and State Curriculum Framework Projects in tandem. It will also look, to the extent we can, at the National Science Foundation's Statewide Systemic Initiative projects. From this study we hope to develop a better understanding of Federal initiatives as they complement or operate independently of each other.

To get more bang from the evaluation buck, we've looked to cooperative efforts across our own evaluation office and with other evaluation offices. Our national evaluation of the Chapter 2 block grant program needed to look at how private school students were participating in Chapter 2—specifically,

what special arrangements were being made for their participation. At the same time, we had commissioned a special study to look at Chapter 1, the categorical program, including how private school students were participating. The solution here was very simple. We decided to piggyback the Chapter 2 items on to the larger Chapter 1 study.

Similarly we're working with the Department of Health and Human Services to examine the impact of the JOBS program on the education of the children of JOBS program participants. To study the linkage to adult literacy, we are pulling funding from adult education evaluation funds.

The national performance review initiative by the Vice President has given us a challenge that I hope we can turn into an opportunity. The Department of Education has volunteered to serve as a reinvention lab. It plans to develop performance indicators for our major programs similar to those being mandated in Public Law 103-62. The staff offices in the department, including our own, are also participating. For our part we are developing, with the help of people like Bob Boruch and the members of the National Academy of Public Administration, ways to look at our own productivity and impact. Bob is helping us by developing a user survey similar to the work described today.

I'm thankful to the National Science Foundation for funding the conference and the work of the authors of the papers presented here. Such conceptional work is rarely undertaken without the prospects of immediate payoffs or knowledge of exactly how the work relates to immediate concerns. NSF is making a valuable contribution to evaluation methodology by leaving these footprints. Other agencies can follow them as they go through the process of thinking how to assess the impact of their work and the programs and projects they support.

It's a pleasure to comment on three such intelligent and creative papers. When I first heard of the concept of footprints, it struck me as being of doubtful usefulness, but I've changed my mind.

I like Johnson's paper primarily because she raises both of the two big questions. One question we all ask in this field is, How do you attribute causes from effects? The question we don't ask often enough is, Compared to What? Programs and explanations compete. Johnson longs for the all-knowing perspective, looking backwards to intention and planning, forward to outcomes, and sideways to what might have been. I think some of this sideways vision is possible, as Bob Yin seems to believe. The field of public policy, a major sponsor of program evaluation, does ask very broad questions about what, in a given era, was on the public agenda; what sorts of efforts were deployed (some nonobvious); and what in the end these led to. Although these questions are not very rigorous, eventually there is historical consensus: Were income maintenance plans cost effective? Was the tax cut of 1981 successful? The logical step here is that what might have happened may have happened. It's helpful when, over a decade or more, streams of evaluation are directed so as to flow down ALL the major channels of program and policy reference, not just the main stream. It makes the historical judgment more complete and more sound.

Let me try to relate this to education. Here are three examples of what are essentially competing explanations for certain broad sets of effects. First, in the cognitive realm, there is an established tradition of work in educational psychology that says that the demonstrated level of achievement in knowledge-item testing, at least a variable portion of the score, is a function of the amount and intensity of specific instruction, of actual brain time-on-task in the delivered curriculum. We in Education and Human Resources would not deny this, but we would think the matter more complicated. The point is that this explanation doesn't concern itself with pedagogy or the quality of thinking by the student or the generativity of knowledge: it talks about measured content exposure, the length of the school year, the sequencing of material and the timing of testing, and so on. If the stated criterion is test score improvement, and program evaluation were to show that this molecular and measurable kind of approach yields interventions that pay off, compared to some of what EHR is doing, it might suggest to those who pay us that some of the stones we are lifting are not worth lifting.

Second, at a higher level of generality, there's a "bet" in the nineties that there is a more powerful avenue for the welfare of young people than educational reform: I refer to the well-child movement involving the integration of human services of all kinds, as pioneered at the Harvard School of Public Health and the Carnegie Councils and funded at quite high levels in the Department of Health and Human Services. An implication here is that the dropout rate in high school is perhaps not fundamentally an instructional matter: to explain it, you need to look at the social aversiveness of schooling for some kids, at the labor market and at foregone earnings for these kids, at the family—including nontraditional families—or the neighborhood or the subculture as an economic enterprise, and at still other kinds of explanations. At any rate, this is the kind of situation where in 20 years experts will say which general strategy was "on target"—although if the identified problem has changed, then the desired target may also have changed.

Finally, there is also a bet going that the tocsin sounded in the early eighties about a competent work force, economic competitiveness, and national security is not really something the schools can solve. The argument, now becoming explicit, is, if business needs an up-to-date technically trained work force,

let them do the training, invest the capital, and capture the benefits; why load it on the schools?

The general point I am making is that, in a medium-long timespan, if we don't want instances of program evaluation to appear at some later time as quaint or irrelevant, we need to keep in mind the definition of the problem and the public choice arena in which a program existed, compared to other problems and choices. That's why I'm pleased with Yin's Exhibit 6, which begins at the right place and ends ... almost at the right place. Legislative mandate refers, inevitably, to some perceived problem or need, where intervention is thought to be possible. With regard to NSF, the Vannevar Bush report and the 1950 enabling legislation refer to a compact between government and especially the military, industry, and universities that would ensure that a domestic Manhattan Project could be mounted at any time of crisis. Later, in a different era, the report language concerning authorization and appropriations for EHR during its rather extraordinary period of budget expansion gives us various statements about why, for what purposes. The corresponding language for the Department of Education presumably has addressed other large issues: the dropout rate, the school-to-work transition, and the problems of multilingualism and multiculturalism. It is important in program evaluation to examine the sense of problems, needs, and possibilities that existed as the program itself came into existence. All I want Yin to do is to bring that analysis around to the right-hand side of the figure, so that we see outcomes with respect to what. That is, what do the new ideas, applications, capacity, and so on address? Is it leading a good life? Is it economic viability at the personal and societal level? Is it raising achievement in school?

This bears directly on Yin's commendable inclusion in his model of two locations for rival hypotheses: that is, competing explanations. The two boxes represent different sorts of processes. The box at the top, subscript 1, refers either to historical convergence of cause to effect processes or to alternative causes or paths to the same effects. That is, these same ideas, applications, influences, and capacities would occur anyhow, for different reasons. In that case, the program in question was in synch with other cause-to-effect processes; at worst, it duplicated them unnecessarily.

The box at the bottom, subscript 2, refers to a narrower kind of explanation: that normal science, including "normal" applied research, is highly overdetermined, reflects the Zeitgeist and runs under its own steam. It is not genuinely *directed* toward the ends shown in the chart, though they may indeed be true consequences. The challenge is that the specific mandate, appropriations, priorities, and funding decisions of an NSF program contributed nothing distinctive: the availability of any orderly decision process would have led to the same quantity and quality of R&D. Examples: there is a technological shift lying behind Research in Teaching and Learning; there is a particular public policy research agenda driving the Studies program.

We are more familiar with this latter kind of "compared to what" challenge in evaluation of granting programs. I have two specific suggestions. First, it is useful to map the portfolio of funded projects onto the set of all fundable research projects: projects designed, proposed, field tested, or conceptualized by a given pool of researchers. If what NSF selects is basically an exact subset of all possibilities out there, across a defined set of research generators, then there is a tight relationship between the field and the program. The field drives the program, the program fuels the field. This is said to be the case in some programs at NIH, where a successful grant-getting investigator always proposes the research he or she has just successfully piloted (or even completed). If the two distributions are not alike, it may be evidence for a specialized ecology, some sort of lock and key fit in research funding: some proposals go to NSF, some to

ED, some to Spencer, and so on. In this case, the differentiated route to outcomes is more easily traced.

I would like Yin's box, Portfolio of Funded Projects, to be shown in relation to another box, called Portfolio of Possible Projects, in some other plane or orientation. This comparison is not done often enough; it is feasible, but it is difficult. As these papers point out, investigators work on different things under the same grant, or on the same thing under different grants, etcetera. Since the outcomes in question are not always measurable in terms of money, it is impossible to construct their production functions in the usual econometric terms. So my second suggestion is to use *time* as the metric. In principle, it is feasible to go into the population of those doing educational research and ask about investments and yields (appropriately discounted) and opportunity costs. Why did you do this research rather than that? When did you expect a payoff? When did it arrive? How much time have you spent not doing research, but volunteering in a high school classroom? Serving on a school board? Lobbying for specific educational practices at the district office or the state house? Teaching a course in the School of Ed—if you're a departmental scientist—or accepting an education graduate student for a dissertation? Urging young faculty to go out into the schools ...? Johnson, in her paper, suggests some of these possibilities, and there have been some useful studies by the Woods Hole circle around Zacharias and Bruner in the early sixties along these lines. After all, researchers choose among research possibilities, and they are not just researchers. If real impacts and outcomes in the educational arena are to be attributed to a full range of causes, or even if the dynamics of the research process are to be fully understood, then these "compared to what" tracings and paths are important.

I apologize to my esteemed friend Bob Boruch for not delving deeply into his paper in this forum. He knows that I think it's full of good ideas. Briefly, I endorse the importance he gives to filter mechanisms and intermediary groups: these are key aspects of both quality control and uptake of information. Overlooked sources of unique information about knowledge into practice include, besides those Boruch mentions, scholarly autobiographies, Festschrifts documenting intellectual circles and institutional histories (e.g., of the Education School at Stanford), and retrospective why-I-worked-on-what-I-worked-on-when-I-worked-on-it volumes such as the one Rossi did for the Russell Sage Foundation a few years ago. And the idea about tapping into the memories of longtime civil servants can be extended to certain retired agency officers, who can give crucial information and advice at important moments without their egos being on the line. (You remember how in John le Carre novels Smiley was always being brought back from retirement or disgrace, because they needed him at Cambridge Circus.)

One thing Boruch just touches on (as does Yin) but which is very important, is that in the grant-giving arena it is impossible to trace effects to causes if the only information used is what the researcher *proposed* to do. All the agencies and most of the foundations do a poor job in documenting what was actually done. Program evaluators are quite familiar with this problem, but it's time we in the agencies took some of the burden off them by doing a better job of record keeping and documentation of first-level outcomes ourselves, that is, what the intervention or activity actually amounted to.

I represent a mission agency, NASA. We are not the National Science Foundation. We are not the Department of Education. Our programs have a specific kind of very results-oriented approach. We have a mission to carry out, and that determines the kinds of programs that we can do.

I was very pleased to discover that, while all the papers described what were called nontraditional research methodologies, I didn't find them nontraditional at all. They all model what should be, and is, good evaluation practice. They are only nontraditional in the sense that they are not often carried out in Federal government work.

One of the things that came through in several of the papers, and which I think is important, is the unit of analysis that should be looked at in evaluating programs. That is, what is the distinction between a program and a project? People often confuse the two. At NASA, for example, we have over 300 different programs, many of which are, in fact, actually small projects. I think each of the papers, in different kinds of ways, encourages us to look at the impact of these projects in the aggregate rather than as individual small effects. Such small projects are going to have a limited impact in that the effects, if not immeasurable, certainly will not be very useful to anyone.

There is also a lot of discussion about the difference between quantitative and qualitative data. I have reflected on this since I have been in the government. Why do we spend so much time and emphasis on the collection of quantitative data about our programs? (How many teachers were served? How many curriculum products have we turned out? and so forth.) I blame that machine—the overhead projector. I feel a little bit vulnerable here because I am not using viewgraphs, and in the government, as well as many other organizations, there is a point where you have to present information about your program that can be summarized on one or two viewgraphs. That almost requires a quantitative approach, so that you can build a little chart with numbers and statistics. I give this challenge to myself, as well as to my colleagues and to the writers of these research proposals: think about creative ways to present the results of research that uses qualitative data and a variety of very creative analyses of all those data. Think about how qualitative information can be summarized and communicated in an effective way so that it really will have an impact on future program operation.

There was not much discussion about needs assessment in the papers, and I think it is very important for all of us evaluators to pay closer attention to that issue. There is a recognition that what drives programs in the Federal government is legislative authorization. But, in many cases, there is a great deal of flexibility. There are options, different choices that can be made about the programs. Those options should be selected on the basis of comprehensive needs assessment, which is almost never done.

Finally, I would like to thank Dr. Boruch for teaching me a new word in his paper, "amanuensis." I was not familiar with that word. For those of you who don't know what it is, it is someone who writes from dictation or copies manuscripts. Very often I feel like this at work, and I think many of my tired colleagues feel the same way. Maybe if we expand our horizons in the production of evaluations our vision will be brightened and our work will become more creative and meaningful.