Conceptual Underpinnings For Program Evaluations Of Major Public Importance: Collaborative Stakeholder Involvement

Zoe A. Barley and Mark Jenness
Western Michigan University

Overview

This paper suggests that three considerations should prevail in the evaluation of National Science Foundation (NSF) programs. First, evaluations of major public significance should provide for a process that gives voice to the key stakeholders of the evaluation. Second, evaluation should be designed and implemented to serve a primary function of program improvement, including enhancing dissemination. Third, NSF program evaluations should be exemplars for individual project evaluations.

Current issues in evaluation that have emerged from a need to develop program evaluations relevant to a wide variety of audiences (stakeholders) are briefly discussed. Additionally, emphasis is placed on the importance of using evaluations to shape programs to enhance effectiveness as they are in progress, rather than on providing post hoc findings that are often not amenable to real world adaptation (dissemination).

As a strategy for evaluation, this paper describes a method of evaluation of multiple projects with common or closely similar outcomes that has been named "cluster evaluation."[1] While aspects of the method can be used retrospectively and could be used to aggregate findings from a program's funded projects, a primary value of cluster evaluation is in the formation—during the course of program activity—of an interactive, collaborative group consisting of project directors, funding agency program staff, evaluators, and other appropriate key stakeholders. Cluster evaluation is, therefore, constructivist in orientation, with the evaluation being constructed out of the shared visions, values, and directions of the cluster group.

Cluster evaluation—a collaborative, project-enhancing, leadership-enabling, outcome and system-focused process—is an appropriate framework for any publicly significant evaluation. Certain elements of this evaluation method may be more directly relevant for overall evaluations of NSF programs than others, and the process can be adjusted to meet the needs of particular situations.

Implications for NSF Programs

In its efforts to identify nontraditional approaches to program evaluation (the "Footprints" project), NSF can learn much from cluster evaluation methodology and its philosophical underpinnings. Diverse multisite programs with common areas of interest seeking to improve overall and individual project efforts and determine effects of program process and accomplishment of outcomes, such as those of NSF, are primary candidates for cluster evaluation. Although cluster evaluation can be applied to many settings in and out of government, for the purposes of this paper, reference will be made to the NSF Research in Teaching and Learning (RTL) program as an example for applying cluster evaluation.

"... evaluations of major public significance should provide for a process that gives voice to the key stakeholders of the evaluation."

[1]This is a new use of the term"cluster evaluation" and bears no resemblance to evaluation forms existing prior to 1988.

"The new evaluator is someone who believes in and is interested in helping programs and organizations succeed."

The RTL program, according to documents supplied to the authors as background for the "Footprints" assignment, "seeks to support new discoveries about how individuals and groups learn, teach, and work more effectively in complex, changing environments." Three important goals of the RTL program are particularly amenable to the use of cluster evaluation: 1) building a coherent and comprehensive base to meet future and current needs of all decision makers, 2) initiating an emphasis on direct teacher and other stakeholder involvement, and 3) helping assure the application of research findings. An interactive, collaborative evaluation process gives voice to front line educators, as well as researchers, in a nonthreatening, practical issues-focused context in which assessment and evaluation become tools to improve practice and to shape programs to serve the full range of interested audiences.

Statement of the Problem—Program Evaluations

Through its "Footprints" project, NSF is exploring alternative, nontraditional approaches to evaluating their efforts, especially those programs focusing on mathematics and science education research and applications of technology. Frechtling, in her introduction to the "Footprints" papers, discusses three concerns about traditional evaluations in the context of NSF needs: 1) given the multiplicity of influences, it is unlikely or impossible that appropriate unidimensional causal statements can be drawn, 2) sole use of quantitative measures are likely to exclude important information, and 3) impact measures, such as student achievement, need to be considered relative to the likelihood of impact in the projects' time frames.

These are important concerns and are discussed in more detail in the context of philosophical underpinnings of cluster evaluation. First, the nature of evaluation data needed, not only for funders but also for a wide array of audiences for the purpose of accountability, project refinement and enhancement, and successful dissemination, is much more complex than previously thought necessary and hence more difficult to obtain.

A second critical issue, however, lies in the purpose of the evaluation itself. Wholey (1983) saw parallels in the public sector use of evaluation to what profit does in the for-profit sector, providing critical feedback that is immediately useful to policymakers and managers. In his 1973 work (Wholey and White) he stated, "the main purpose for evaluation, . . . to feed back information about how a program is working to improve its operation, is missing from most local and state evaluation activities." In another article, he suggested,

> The new evaluator is a program advocate—not an advocate in the sense of an ideologue willing to manipulate data and to alter findings to secure next year's funding. The new evaluator is someone who believes in and is interested in helping programs and organizations succeed. At times, the program advocate evaluator will play the traditional critic role; challenging basic program assumption, reporting lackluster performance, or identifying inefficiencies. The difference, however, is that criticism is not the end of performance-oriented evaluation; rather it is part of a

> larger process of program and organizational improvement (Bellavita, Wholey, and Abramson 1986, p. 289).

Finally, Frechtling notes recent trends in evaluation which seek to involve all the stakeholders in the process. Cronbach and associates (1980) see this as a key task in understanding the political nature of the end result of any important evaluation study. Guba and Lincoln (1989) speak of stakeholders' claims, concerns, and issues as organizers for the evaluation. Donmoyer (1991) strongly suggests that stakeholders be actively involved in dialogs before, during, and after the evaluation.

If these purposes and intents—implementing a more appropriately complex evaluation, shaping programming and improving projects by providing feedback during project implementation, and involving stakeholders in the process—are valid, the evaluations should be shaped "upfront" with these goals in mind. Grantees, however, are often not prepared with either the evaluation skills required or knowledge of the broader context in which their project findings are relevant for those findings, to be meaningful. While some amount of information can be obtained from evaluations after the conclusion of projects, to achieve a measure of information appropriate for use in guiding selection of new projects, in disseminating results to other project sites, or for use in systemic change modalities, the evaluation process must be improved as the projects proceed.

Design Considerations—Undergirding Philosophy

Two conceptual models offer useful insights for designing nontraditional evaluations for NSF research-oriented programs: Cronbach's concept of a Social Problem Study Group from his 1980 book, *Toward Reform of Program Evaluation*, and Guba and Lincoln's fourth generation evaluation from the book (1989) by the same title. They also provide guidance in the design and implementation of cluster evaluation described in a later section.

Cronbach suggests the formation of a social problem study group made up of members representing all concerned parties for evaluations of social significance, not unlike panels NSF convenes for evaluation purposes. The group, however, would embrace the following activities:

- Study problems (e.g., What should be the influence and direction of an NSF program?) in the broadest possible way.
- Hear from those who conduct evaluations, preferably as their work progresses; hear from those who deal with the problem in service agencies; hear from those who have ideas about new policies and interventions.
- Produce a far more comprehensive and dependable interpretation than emerges from a single study or a lone critic questioning a finding.
- Continually reformulate the questions worth studying and recast key terms that define stated problems.
- Put research into proper time perspective, dispelling the illusion that quick and partial studies will resolve dilemmas.
- Provide a forum for putting observations and uncertainties into perspective.

"Collaboratively, they generate an evaluation that is far more than monitoring or accountability, but which addresses broad-level policy considerations in a future-oriented mode."

- Be willing, and able to think hard about the specified problems.

In another but related direction, Guba and Lincoln have defined "fourth generation evaluation" in which the processes of the evaluation are as follows:

1. Identifying the full array of stakeholders who are at risk in the projected evaluation.
2. Eliciting from each stakeholder group their constructions about the evaluation and the range of claims, concerns, and issues they wish to raise in relation to it.
3. Providing a context and a methodology through which different constructions, and different claims, concerns, and issues, can be understood, critiqued, and taken into account.
4. Generating consensus with respect to as many constructions, and their related claims, concerns, and issues, as possible.
5. Preparing an agenda for negotiation on items about which there is no, or incomplete, consensus.
6. Collecting and providing the information called for in the agenda for negotiation.
7. Establishing and mediating a forum of stakeholder representatives in which negotiation can take place.
8. Developing a report, probably several reports, that communicate to each stakeholder group any consensus on construction and any resolutions regarding the claims, concerns, and issues they have raised.
9. Recycling the evaluation once again to take up still unresolved constructions and their attendant claims, concerns, and issues (Guba and Lincoln 1989).

Taken together these two frameworks suggest an evaluation process that actively involves all the known stakeholders. Collaboratively, they generate an evaluation that is far more than monitoring or accountability, but which addresses broad-level policy considerations in a future-oriented mode.

Evaluation Questions for NSF Programs

The following questions are suggested as the guiding overarching questions for evaluating NSF mathematics and science education programs, including the Research in Teaching and Learning program. Additional overarching questions and/or subquestions pertinent to a particular program area should be added as appropriate.

The use of concise questions in each of three areas—outcomes, context, and implementation—provides the perspective for not only reporting results, but also for understanding the conditions in which the results were obtained and the exact nature of the programming that produced the results, or lack thereof.

Outcome Questions

What has been the nature of the impact (intended and unintended) of the program on teachers and learners? Positive outcomes? Negative?

What has been the nature of the impact on the system of mathematics and science teaching and learning?

What kinds (and numbers) of new leadership have emerged within the educational system as a result of the program?

What new national or local programs and policies have emerged or been furthered as a result of the program?

Context Questions

Has the program effectively served a diverse body of mathematics and science educators?

Has the program effectively reached a broad range of mathematics and science learners?

For what educational settings has the program's effectiveness been demonstrated?

Has the program funded grantees across a broad range of characteristics representative of the educational system, especially in mathematics and science?

Implementation Questions

Has the program been effective in selecting grantees within categories best able to provide practice-relevant findings?

Have grantees been encouraged and supported to maximize project success?

In understanding project effectiveness, have teachers and learners had a voice?

Is the program sensitive to and implementing projects that result in disseminatable findings?

In sum, the questions should cover not only what has been accomplished within the program but for whom those accomplishments apply and under what conditions. If the answers to the questions are derived through a collaborative process engaging representatives of the various audiences in a consensus-building process, the results are more likely to be applicable to the educational system and not fragmentally to one or another part of the system.

One Strategy for Collaborative Evaluation—A Brief Description of Cluster Evaluation

What follows is a description of an evaluation method that engages a group—or cluster—of projects in common evaluation efforts. Using this method, the authors have been able to accomplish the purposes discussed earlier for NSF evaluation. Cluster evaluation provides a complex, rich data set, derived to a large extent from the involvement of stakeholders in the formation of the evaluation itself, that provides information for determining program impact, as well as improving programs. The process of the cluster also enables and prepares project directors to improve their own evaluation skills and allows them to be better consumers of evaluation data. The authors believe the cluster evaluation model has widespread application in the NSF arena.

The generic method of cluster evaluation was described and named by the W.K. Kellogg Foundation and is used in their various funding initiatives. Implementation, however, varies from cluster to cluster. The specific cluster evaluation method developed and used by the authors with two groups of 12 science education projects is summarized below.

"... cluster evaluation is a complex process with diverse components requiring a variety of skills and resources ..."

Organizing the Cluster

The specific organization of the cluster is affected by several factors, including the number of projects funded, geographic location of projects, nature of topical area, targeted populations, and degree of similarity of the project implementations. Availability and level of experience of the cluster evaluators also affects the process.

Selection of cluster evaluators is initiated by the funder, and basic organization, time frame, role of funder program staff, evaluators, and project staff, and implementation procedures for the cluster evaluation are negotiated.

Projects selected for inclusion in the cluster are usually determined by the funder. Completion of selection of projects varies, with some selected prior to the initiation of the evaluation and others selected several months into the process. Based on the authors' experiences, selection prior to initiation of the cluster evaluation results in a more effective evaluation.

The number of projects in a cluster can vary, depending on the factors described above. The basic purpose and expected results of the cluster evaluation should be carefully considered, along with available financial and other resources. Clusters of not more than 25 are optimal for conducting an intensive collaborative cluster evaluation as described in this paper.

Regular, networking conferences are organized by cluster evaluators and program staff, with funding included in cluster evaluator budgets or a separate budget. Additionally, resources must be made available to funder-program staff to participate in conferences and technical assistance.

A retrospective cluster evaluation of completed projects is also possible, but would necessitate assembling directors from completed projects. The purpose and results of a retrospective cluster evaluation would be different from one with a formative emphasis.

NSF research-oriented projects, such as those in the RTL program, could be easily placed in clusters based on a set of factors, from topic to implementation strategy, and determined by specified evaluation purposes. A retrospective cluster could be determined by regional or other representative sampling techniques.

Cluster Evaluation Team

Because cluster evaluation is a complex process with diverse components requiring a variety of skills and resources, a team of evaluators should be enlisted. It should include people with evaluation expertise, research skills, human relations skills (including writing skills), and appropriate content-area knowledge. Additionally, adequate support staff must be available to attend to details of networking conferences, data collection/compilation, communications, etc. Although not all team members necessarily have to devote full time to the effort, sufficient professional staff time must be available to coordinate and carry out the many evaluative tasks.

In the case of the science education cluster evaluations conducted by the authors, the cluster evaluation team is made up of two principal investigators, one with a strong background in research and evaluation, the other with extensive experience in science education. Additionally, doctoral students and staff bring research, evaluation, organizational, and communication skills to the team. Keeping current in the content area is

necessary if evaluators are to provide useful information to improve programs and to judge outcome accomplishment.

Additionally, external content area and evaluation specialists should be enlisted to periodically review the cluster evaluation.

Setting Expectations

It is important to set expectations for the cluster evaluation up front not only for funders and cluster evaluators, but also for project directors and their staff. Although some projects may have a proposed evaluation plan, including an internal evaluator to implement it, most will need assistance with both internal and cluster evaluation activities. Expectations for funded projects must include full participation in all cluster evaluation activities, including networking conferences, data collection and analysis, and reporting and dissemination. Funders must make these expectations clear and provide adequate resources to facilitate full participation.

Through RFPs or in award letters, NSF staff would make expectations clear for full participation in the cluster evaluation. Additional communications would introduce cluster evaluators and provide instructions for collaboration.

Negotiated Common Cluster Outcomes

Usually at the first networking conference following selection of projects for the cluster, initial common cluster outcomes are determined collaboratively. Using important evaluation questions developed by project and funder program staff for specific projects and questions developed by cluster evaluators and program staff for the overall cluster, a comprehensive list of outcomes is devised.

From this list, a set of common cluster-level outcomes is developed by consensus of the project directors and evaluators, funder program staff, and cluster evaluators. In one science education cluster, 19 cluster outcomes, held in common by two or more projects, were created addressing issues related to students, teachers, curriculum, collaboration, and continuation/ dissemination.

As projects evolve and the cluster evaluation develops, modifications are made to the common cluster outcomes as appropriate, such as adding outcomes or modifying existing ones to better reflect actual intended outcomes. This set of outcomes provides a partial framework for the evaluation of the cluster of projects, and "represents to the projects the intended impact of the cluster" (Barley, 1991). Individual project-level evaluations may also be conducted by projects in the context of the cluster evaluation, depending on requirements of the funder.

"As projects evolve and the cluster evaluation develops, modifications are made to the common cluster outcomes as appropriate..."

For use at NSF, staff, along with cluster evaluators, would develop a set of important questions for the overall evaluation of, for example, a cluster of RTL projects. Some questions will be pertinent to the overall RTL program and others specific to the particular cluster of RTL projects. Individual project staff develop important questions pertinent to their own projects. Collaboratively, a set of common cluster outcomes is then established through negotiation.

Collaborative Data Collection

Both qualitative and quantitative data come from a variety of sources and are in a variety of forms. Individual projects collect data directly from the participants through questionnaires, interviews, observations, journals, standardized tests, recordkeeping, and common

cluster instruments (same instruments used across projects to collect consistent data). Some data are reported in annual reports; other data are sent directly to cluster evaluators. Cluster evaluators collect data from cross-project participant surveys, project staff interviews, documents, participant interviews, and site visits and observations. Also collected is specific information on the strategies and activities each project uses to accomplish the cluster outcomes, as well as contextual information pertinent to cluster outcomes.

"When expectations for data collection are clear early in the process, ... better data are the result."

Several factors affect the quality and quantity of data, including commitment of the various stakeholders to the process, financial resources, and data collection design. When expectations for data collection are clear early in the process, and cluster evaluators facilitate the process through technical assistance and instrument development, better data are the result.

It would be important for projects within an NSF cluster to collect data pertinent to individual project and cluster outcomes, as well as contextual factors and implementation strategies. With technical assistance from cluster evaluators, project directors and their staff will be in the best position to collect pertinent data for individual project and cluster use. Cluster evaluators would also conduct across-project data collection efforts.

Regular Networking Conferences

Direct networking among all project directors, project staff and evaluators, cluster evaluators, funder program staff, and guests at annual or semi-annual networking conferences is an important component of cluster evaluation. The purposes of these conferences will vary somewhat depending on the purpose of the evaluation, topical focus of the cluster, and frequency of the meetings. All networking conferences should include sessions (1) to conduct strategic planning for, exchange ideas about, provide direction to, discuss issues and problems emerging from, and review and analyze data and findings of the cluster evaluation; (2) share lessons learned with other projects; and (3) visit project sites. For a science/mathematics education focused cluster, for example, purposes should also include learning about current and developing issues in science education and science curriculum, instruction, and assessment topics directly pertinent to projects; and formally and informally sharing science education curriculum materials and instructional strategies. Networking is at the heart of a constructivist approach, since it provides a forum for direct engagement of major stakeholders in the cluster evaluation process.

Networking conferences are organized collaboratively between cluster evaluators, program staff, and project directors. Specific travel, overnight accommodation, meal, and meeting arrangements can be part of the cluster evaluator's responsibility and funded accordingly, or the funder can arrange or contract for networking conferences. The number and duration of conferences are related to the purpose of the cluster evaluation and/or available financial resources.

For a cluster evaluation of NSF projects, program staff would be actively involved with cluster evaluators in planning and implementing the conferences. Project directors, individually and in committees, provide feedback and can help make arrangements for the gatherings.

Darn Analysis and Working Hypotheses

A method used in one of the authors' science education clusters to review and analyze the diverse outcome-related data is the use of "working hypotheses," a term first coined by Cronbach (1975), describing tentative hypothesizing statements "that give proper weight to local contextual conditions," but facilitate transferability across varying contextual situations. The degree of transferability depends on the similarity between contexts—the "fittingness" or "degree of concurrence between sending and receiving contexts" (Lincoln and Guba, 1985). Core and auxiliary working hypotheses, based on common cluster outcomes, address commonalties and differences in project-level implementation strategies (Barley, 1991). Working hypotheses are reviewed and modified at networking meetings. Tentative findings are developed by the evaluators and presented to the cluster members for further review. Project staff have an opportunity to offer suggestions for modifications based on additional data and findings from individual projects and make recommendations for additional relevant data collection.

Other analysis methods for mixed data can be used, but should involve project directors and staff at appropriate points in the process.

Cooperative Derivation and Dissemination of Results

Dissemination of findings and sharing of lessons learned occurs between individual projects in the cluster, from individual projects to other pertinent programs (for example, science/mathematics education programs for an NSF cluster), among projects at networking conferences, and at local, state, and national gatherings of educators, evaluators, and others. Networking conference sessions are also devoted to planning common dissemination activities, such as development of printed materials, videos, conferences, consulting services, etc.

This will be an important aspect of cluster evaluation for NSF programs, since project directors and staff must be actively involved in deriving and disseminating results, not only of the evaluation, but of project research findings. Evaluation findings should help NSF program staff, in collaboration with cluster evaluators and project directors, determine future funding and research efforts. Networking within a cluster and between clusters would also facilitate interactions among a large group of researchers and NSF staff, leading to more informed coordination of NSF-funded research activities and their relationship to overall education reform efforts.

"... cluster evaluation engages stakeholders in the evaluation process."

Recommendations and Conclusions

Cluster evaluation as briefly described in this paper is an innovative and effective method that can be appropriately adapted to help meet the needs of the National Science Foundation as it seeks to develop an evaluation framework that will identify the footprints left behind by its programming efforts. Although cluster evaluation can be used retrospectively, it is particularly appropriate when used with groups of projects initiating and conducting their programs, thus identifying footprints throughout the course of the projects.

As a formative/summative combination approach (as described in this paper), cluster evaluation engages stakeholders in the evaluation process. It provides feedback to projects as they implement their programs, and, thus, helps them improve. Cluster evaluation also

measures the overall impact of the group of projects and addresses contextual factors and implementation strategies.

Using it retrospectively, cluster evaluation provides a framework for addressing important evaluation questions related to outcomes, context, and implementation. It is suggested that an evaluation "panel," representative of a broad cross-section of NSF stakeholders, project directors, program staff, evaluators, teachers, and learners, be established for particular NSF program areas or portions of a program area (i.e., projects with similar missions). Operating collaboratively and on an ongoing basis, their purpose would be to construct and adjust the evaluation design out of their shared concerns, values, and directions for the program. They would jointly establish the evaluation questions, determine the specific design, collect common data, and develop analyses appropriate to the real world of educational practice.

References

Barley, Z.A. 1991. Strengthening community-based project evaluations and deriving cross-project findings. Paper presented at the meeting of the American Evaluation Association, October 1991, Chicago.

Bellavita, C., Wholey, J.S., and Abramson, M.A. 1986. Performance-oriented evaluation: Prospects for the future. In *Performance and Credibility: Developing Excellence in Public and Non-Profit Organizations*, J.S. Wholey, M.A. Abramson, and C. Bellavita, (eds.). Lexington, MA: Lexington.

Cronbach, L.J. 1975. Beyond the two disciplines of scientific psychology. *American Psychology* 30(2): 116-27.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., and Weiner, S.S. 1980. *Toward reform of program evaluation*. San Francisco: Jossey Bass.

Donmoyer, R. 1991. Postpositivist evaluation: Give me a for instance. *Educational Administration Quarterly* 27(3): 265-96.

Guba, E.G., and Lincoln, Y. S. 1989. *Fourth generation evaluation*. Newbury Park, CA: Sage.

Lincoln, Y.S., and Guba, E.G. 1985. *Naturalistic inquiry*. Newbury Park, CA: Sage.

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

Wholey, J.S., and White, B.F. 1973. Evaluation's impact on Title I elementary and secondary education program management. *Evaluation* 1: 73-76.