New Directions for Evaluation Proposal
May, 2005

I. Proposed Title of Volume:
New Perspectives on the Evaluation of Science, Technology, Engineering and Mathematics

II. Guest Editors:

Douglas Huffman, Ph.D. is an Associate Professor of Science Education at the University of Kansas with a specialty in program evaluation. He has worked in the field of program evaluation for over 15 years, and during this time has served as a program evaluator on numerous National Science Foundation funded projects. Most recently he has evaluated the NSF funded State Systemic Initiatives (SSI), and the Collaboratives for Excellence in Science and Mathematics Teaching (CETP). He is currently the PI on an NSF funded project called Collaborative Evaluation Communities in Urban Schools. He has published his work in both science education and evaluation journals. His work appears in the Journal of Research in Science Teaching, School Science and Mathematics, The American Journal of Evaluation, and The Canadian Journal of Program Evaluation.

Douglas Huffman, Ph.D.
Associate Professor – Science Education
University of Kansas
Department of Teaching & Leadership
1122 West Campus Road
Lawrence, KS 66045
(785) 864-9675
Huffman@ku.edu

Frances Lawrenz, Ph.D. is a professor in the Department of Educational Psychology at the University of Minnesota. She has received College of Education and Human Development recognition as the Wallace Professor of Teaching and Learning and the University wide award for outstanding contributions to graduate education. She also has administrative experience as department chair, Associate Dean of the College and as the Assistant Vice President for Research and Associate Dean of the Graduate School of the University. Her specialization is science and mathematics program evaluation and she has been working in that field for 35 years. She has conducted numerous evaluations of NSF projects and programs and has twice served at NSF in a rotator position. She is currently the PI or lead evaluator for twelve NSF projects and has numerous publications in evaluation and science and mathematics education journals.

Frances Lawrenz, Ph.D.
Professor – Educational Psychology
University of Minnesota
Department of Educational Psychology
178 Pillsbury Drive SE
Minneapolis, MN 55455
(612) 625-2046
Lawrenz@umn.edu
III. Overview of Volume:
This proposed volume includes state-of-the-art viewpoints, examples, and methodological approaches to the evaluation of science, technology, engineering and mathematics (STEM). STEM evaluation has always been an important area in the field of evaluation given the issues facing public schools today and the economic and social considerations of STEM fields. However, in recent years there has been renewed interest on the methodological issues of STEM evaluation due to the focus of the federal government on “scientific” evaluation, and attempts to promote experimental designs to better attribute causation to program activities. In this issue we will describe new approaches to STEM evaluation that align with more contemporary views of “scientific” evaluation. An issue of New Directions in Evaluation on STEM evaluation will serve as an excellent place to ground the discussions of “scientific” evaluation and will have broad applicability to the field of evaluation as a whole through discussions of methodological and theoretical issues of evaluation.

IV. Justification:

Scope: A volume of New Direction for Evaluation on using the evaluation of Science, Technology, Engineering and Mathematics (STEM) to highlight and discuss the implications of scientific evaluation is important because of the renewed focus on “scientific” approaches to evaluation. STEM evaluation has traditionally used many different methods and philosophical approaches to evaluate programs; however, recently there has been a political movement to focus on experimental approaches to evaluation. While this movement affects all fields of evaluation, the movement acutely affects STEM evaluation because stakeholders in STEM fields traditionally have more scientific backgrounds and philosophies. As part of this movement the United States Department of Education placed a priority on “scientifically” based evaluation methods (U.S. Department of Education, 2003). The priority advocates particular evaluation methods over others with the belief that “scientific” approaches are preferred. Specifically, there is a belief that evaluation methods using more experimental design are best for determining project effectiveness. Historically, the desire for scientific evidence is rooted in the U.S. Government Performance and Results Act (GPRA), passed in 1993, which requires federal agencies to determine measurable goals for all of their program activities. Agencies must measure their performance against goals and report progress to Congress as part of the annual budget submission.

Against this background, the national commitment to accountability has expanded. This began with the development of standards for education such as those developed by the fields of science and mathematics (National Council of Teachers of Mathematics, 1989, 1991, 2000; National Research Council, 1996), and was followed by the passage of the “No Child Left Behind” Act (2001). In fact, the Elementary and Secondary Education Act (ESEA) as reauthorized by the No Child Left Behind Act of 2001 (NCLB) used the term “scientifically-based” more than 100 times in the context of evaluating programs to determine their effectiveness. State departments of education across the United States are also engaged in the most comprehensive state testing this country has ever seen in an effort to satisfy the accountability requirements in the No Child Left Behind Act. In 2002, the Office of Educational Research and Improvement was reconstituted into the Institute of Education Sciences reflecting the intent of the federal government to advance
the field of education by making it more rigorous and supporting more evidence-based education. All of these events point to a significant national movement towards using scientifically based evidence to evaluate programs. Such a movement has important implications for the field of evaluation. It is important for evaluators to understand the underlying assumptions of such a priority and the implications for practice.

The priority on “scientific” evaluation raises many questions about what we mean by “scientific” evaluation and what are the best ways to evaluate programs. In recent years there have been advances in evaluating STEM programs that could help the field as a whole address the issue of “scientific” evaluation. There have been advances in mixed method approaches, multi-site participatory evaluation, cultural competency, and issues of social justice. All of these perspectives go beyond the current traditional experimental view of scientific evaluation and expand the notion of what it means to evaluate STEM programs. In this volume, we will present new methodological approaches to STEM evaluation and advance new notions of evaluation for science, technology, engineering and mathematics. We will also include state-of-the-art examples of STEM evaluation from some of the leading experts in the field that will help readers see how these new approaches to STEM evaluation can be applied in practice.

**Novelty and Timeliness:** This new priority on “scientific” evaluation has major implications for the entire field of evaluation and for the field of STEM evaluation specifically. While this priority is currently coming from the U.S. Department of Education, it appears to be influencing other federally funded programs and private funders alike. This in turn has the potential to significantly affect the way program evaluation is conducted on federally funded social service programs. It is a timely issue that has significant implications for the field of evaluation.

Related publications on this issue have emerged in response to the National Research Council’s Committee on Scientific Principles for Education Research, Scientific Research in Education, (National Research Council, 2002). This report emphasized the importance of rigorous scientific methods. The report stimulated a debate, which continues today, about how to conduct scientific inquiry. The new priorities for scientific research have generated much debate among educational researchers. Numerous articles have been written on the implications of this directive and concerns over the view of science put forth by the U.S. Department of Education (Jacob & White, 2002; Eisenhart & Towne, 2003). The debate has centered on arguments regarding definitions of “scientifically-based research” and whether or not such priorities limit educational research rather than improve it. Feuer, Towne & Shavelson (2002) and others (Jacob & White, 2002) argue that emphasis should be placed on understanding and appreciating multiple perspectives rather than advocating particular methods.

The debate on education research directly relates to the debate on evaluation. The same issues of what constitutes “scientific” apply. The field of evaluation is in need of research on which particular methods are best in which situations. To state that experimental methods are “best” for evaluating program effectiveness fails to consider our understanding regarding the effectiveness of different methods. There have been numerous methodological advances and new perspectives in STEM evaluation that would help the field see how one can conduct “scientific” evaluation while at the same time moving beyond the traditional experimental view of evaluation.
Quality: The proposed volume on STEM evaluation will include state-of-the-art work in program evaluation from some of the top STEM evaluators in the United States. We have agreements from top experts in the field of STEM evaluation to write manuscripts on important methodological and theoretical issues in the field. They will also provide examples from their own practice on how to apply these ideas in practice. In addition we will provide a critical analysis in the volume by including experts who have different opinions on how to engage in STEM evaluation. Dr. Frances Lawrenz, University of Minnesota will contribute a manuscript on mixed-method approaches to STEM evaluation. This manuscript will be based upon her work evaluating STEM programs using a variety of approaches. Dr. Douglas Huffman, University of Kansas will contribute a manuscript on evaluation capacity building in STEM education. This manuscript will include a new model of evaluation capacity building that is currently being used as part of a National Science Foundation grant to develop collaborative evaluation communities in urban schools. Dr. Arlen Gullickson, Director of the Evaluation Center at Western Michigan University will write a manuscript on large scale, multi-site STEM evaluation. We will also include a manuscript from Dr. Conrad Katzenmeyer, former NSF program officer in the Directorate for Research, Evaluation and Communication. He will provide a federal perspective on the challenges facing STEM evaluation. Dr. Katzenmeyer has served at the National Science Foundation for several decades and can provide a historical perspective on the new directions of STEM evaluation. To provide a more critical perspective on STEM evaluation we will include contributions from Dr. Jennifer Greene, University of Illinois and Dr. Donna Mertens, Gallaudet University, who will encourage the field to consider how issues of social justice and cultural competency can also be considered in STEM evaluation.

Audience: The proposed volume on STEM evaluation will be written for a diverse audience including both STEM and non-STEM evaluators. The goal is to produce a volume that will have interest for the large number of readers who engage in evaluation. This includes individuals who are involved in the evaluation of science, technology, engineering and mathematics evaluation. In addition, we will focus on methodological issues of multisite evaluation, participatory evaluation, and mixed method approaches. Together these two focus areas on STEM and methodological and theoretical issues of what constitutes “scientific” evaluation will have fairly broad appeal to the readers of NDE.

Level of Presentation: The proposed volume does not deal with technical materials that would not be familiar to the wider evaluation audience. The methodological and theoretical issues will have broad understanding to the field of evaluation as a whole. Specific STEM evaluation examples will be explained such that readers from other fields can understand STEM issues.

V. Outline of Volume

1) Critical Issues in STEM Evaluation – Douglas Huffman, Frances Lawrenz & Kelli Thomas

2) Mixed Method STEM evaluation – Frances Lawrenz & Douglas Huffman

3) Local Evaluation in Multisite STEM Programs: Relating Evaluation Use and Program Results – Arlen Gullickson & Carl Hanssen
4) National Science Foundation Perspectives on the Nature of STEM Program Evaluation – Conrad Katzenmeyer

5) Cultural Competency in STEM evaluation – Donna Mertens & Rodney Hopson


7) A Model of Evaluation Capacity Building in STEM Education – Douglas Huffman, Frances Lawrenz, Kelli Thomas & Lesa Clarkson

8) Synthesis of Ideas – Frances Lawrenz & Douglas Huffman

VI. References:


This manuscript will set the stage for the key issues in STEM evaluation. The manuscript will provide readers with an historical perspective on STEM evaluation and will discuss the debate regarding scientific approaches to STEM evaluation. Traditionally, STEM evaluation has used many different methods and philosophical approaches to evaluate program; however, recently there has been a political movement to focus on traditional experimental approaches to evaluation. While this movement affects all fields of evaluation, the movement acutely affects STEM evaluation because stakeholders in STEM fields traditionally have more scientific backgrounds and philosophies. As part of this movement the United States Department of Education placed a priority on what are referred to as “scientifically based evaluation methods.” (U.S. Department of Education, 2003). Previously, the emphasis on scientifically based evidence has focused on educational research, but with this new priority aimed at the field of evaluation it now appears that the stakes have been raised for evaluators. The priority places more value on projects employing an evaluation plan that is based on experimental methods. The priority suggests that particular evaluation methods are better than others. Specifically,

Evaluation methods using an experimental design are best for determining project effectiveness. Thus, the project should use an experimental design under which participants—e.g., students, teachers, classrooms, or schools — are randomly assigned to participate in the activities being evaluated or to a control group that does not participate in the project activities being evaluated. (p. 2)
Next on the priority list of evaluation methods are quasi-experimental designs with carefully matched comparison conditions, which may be used when random assignment is not possible. In cases where an extended series of observations of the outcome of interest precedes and follows the introduction of a new program or practice, regression discontinuity designs may be employed. When sufficient numbers of participants are not available, single-subject designs such as multiple-baseline or treatment reversal or interrupted times series may be used if they are capable of demonstrating casual relationships. This new priority has major implications for the field of evaluation as a whole. While this priority is currently coming from the U.S. Department of Education, it is having implications on other federally funded programs and on private funders alike. This in turn affects the way program evaluation is conducted. In this article, we will argue that the view of scientific evaluation put forth by the Department of Education uses an overly-simplistic view of the nature of science and risks limiting the evaluation of programs rather than improving them as intended.

Historically the desire for rigorous scientific evidence is rooted in the U.S. Government Performance and Results Act (GPRA), passed in 1993, which requires federal agencies to determine measurable goals for all of their program activities. Agencies must measure their performance against goals and report progress to Congress as part of the annual budget submission. Against this background, the national commitment to accountability has expanded. This began with the development of standards for education such as those developed by the fields of science and mathematics (National Council of Teachers of Mathematics, 1989, 1991, 2000; National Research Council, 1996), and was followed by the passage of the “No Child Left Behind” Act (2001). The Elementary and Secondary Education Act (ESEA) as reauthorized by the No Child Left Behind Act of 2001 (NCLB) used the term “scientifically-based” more than 100 times in the context of evaluating programs to determine their effectiveness. State departments of education across the United States are also engaged in the most comprehensive state testing this country has ever seen in an effort to satisfy the accountability requirements in the No Child Left Behind Act. In 2002, the Office of Educational Research and Improvement was reconstituted into the Institute of Education Sciences reflecting the intent of the federal government to advance the field of education by making it more “scientific” and supporting more evidence-based education. However, it has become increasingly clear that the view of science put forth by the Institute of Education Sciences is actually quite narrow and appears to equate scientific with experimental.

The belief in the value of experiments over other types of evidence is based upon the materialistic objective Western world-view that believes an experiment provides objective evidence to answer questions. The Institute of Education Sciences capitalizes on this view and refers to this view as “scientific evaluation” when what they seem to be referring to are medical trials and experiments. They do not place as much value on other scientific methods, such as field studies or observations. Some have argued an experimental view of science is overly simplistic (Berliner, 2002) and does not represent a realistic and comprehensive view of science. (Feuer, Towne, & Shavelson, 2002; Jacobs & White, 2002). Science is certainly a discipline with a long history of using experimental methods; however experiments are only one of many methods that are considered “scientific.” In reality science relies upon a wide variety of methods. Observation methods, descriptive methods, and non-experimental methods are all methods of conducting scientific research. For example, astronomers very rarely employ
experimental methods. Observational methods are a significant part of the field of astronomy
and are regarded as the most appropriate methods. Likewise for ecologists or field biologists.
Much of our understanding regarding animal behavior and the complex relationships in
ecosystems has relied upon observational methods, not experiments. According to the new
priority such methods would not be preferred and would not be as highly valued as experimental
methods even though they are the most appropriate methods for the particular field and to answer
questions. The over reliance on experiments as a means to determine cause and effect, or to
determine if a program is effective, fails to take into account that other methods are not only
equally important, but can lead to a better understanding. If one only relies upon experiments to
evaluate program one would be left with an incomplete understanding of programs. There are
many questions about programs that cannot be answered with an experiment, and to fully
evaluate programs it is essential to employ a wide variety of methods.

It is also interesting to note that the U.S. Department of Education advocates specific evaluation
methods given there is so little “scientific research” to support specific methods of evaluation.
As Henry and Marks (2003) point out, publications on how to do evaluation are “generally based
on personal experience, observation, and the individual’s sometimes idiosyncratic beliefs and
values – and not on carefully gathered evidence that can be described, shared, and critiqued.” (p.
70) The field of evaluation is in need of research on which methods are best in which situations.
To state that experimental methods are best for evaluating program effectiveness fails to consider
our understanding regarding the effectiveness of different methods. Stufflebeam (2001)
categorized over twenty different evaluation approaches and assessed their strengths and
weaknesses, and considered whether, when, and how the methods are best applied. Carefully
considering a wide variety of evaluation methods seems like a much more fruitful approach than
merely advocating a particulate method or set of methods. Overall, this manuscript will help to
set the stage for the following manuscripts on STEM evaluation. The over-arching theme will be
that scientific evaluation is a broad comprehensive concept that includes both experimental and
non-experimental methods, and that a wide variety of methods are preferred for STEM
evaluation and for the field of evaluation as a whole.

References

There is continuing debate regarding what constitutes appropriate methods for conducting evaluations. The U.S. Institute of Educational Sciences has advocated for a “gold standard” of randomized controlled experimentation; however, not everyone agrees that the standard is “gold.” Maxwell (2004) points out that limiting evaluation to “scientific” approaches is tantamount to assuming only the regularity view of causation and forcing an emphasis on determining what “happened” rather than considering how it did so. In other words, it de-emphasizes the importance of understanding process and meaning which is critical to effective program evaluation. Feuer, Towne, and Shavelson (2002) and others (Jacob & White, 2002; Eisenhart & Towne, 2003) argue that emphasis should be placed on understanding and appreciating multiple perspectives rather than advocating particular methods. Recently the National Academies sponsored a colloquium to discuss research methods. In a paper prepared for that colloquium, Raudenbush (2004) made the often-stated compromise that the methodology should match the question rather than proposing a single methodology. By suggesting this seemingly obvious compromise, Raudenbush and others are sidestepping the real issue. Specifying a single methodology for evaluation limits the types of questions that can be investigated and ultimately allows only one system of values. House (2004) suggests that focusing the debate in the evaluation field on methods rather than on beliefs or values obscures the philosophical differences underlying the different value systems. As Burbules (2004) points out, it is critical to bring these underlying differences out into the open so that constructive debate can take place and evaluation can be optimized.

There has been increasing interest in mixed methods. Mixing methods is another effective way to address the issue of preferred evaluation methods. Chatterji (2004) presents an argument for extended term mixed method evaluation designs. She suggests that formal study of contextual and site-specific variables with multiple research methods is a necessary prerequisite to designing sound field experiments for making generalized causal inferences. Johnson and Onwuegbuzie (2004) suggest that mixed methods is a research paradigm whose time has come. They posit that mixed methods research is aligned with pragmatism and proposed an eight-step process for conducting mixed methods research. They claim that the methodological pluralism of mixed methods could result in superior research. Earlier, Greene and Caracelli (1997) provided a much more detailed examination and categorization of mixed methods evaluation where they discussed how methods might be combined most effectively. They also suggested that methods
are embedded within philosophies and represent paradigms as opposed to simple methods. Furthermore, Caracelli and Greene (1997) suggested two different mixed methods design categories: component designs that include triangulation, complementarity, and expansion; and integrated designs that include iterative, embedded, holistic, and transformative. A component design is one in which different methods are used but are not integrated with each other. A component triangulation design is when different methods are used to assess the same phenomenon in order to provide convergence of opinion and increased validity. A component, expansion design is when the different methods are used to assess different aspects of a program. An integrated design is when the different methods are used in ways that require they interact with each other. Iterative, integrated design is when one type of method is used and then another; embedded is when one method is used within a broader look utilizing a different method; holistic is when the results from the different methods are combined to provide an overall perception of the evaluation object; and transformative is when the methods are used to move the evaluation above any one method. These different approaches can be used in a variety of ways to enhance evaluation. Lawrenz and Huffman (2002) suggest a more holistic approach to mixing methods and propose the metaphor of an archipelago in discussing how the results of different methods might be combined.

Unfortunately many evaluations are designed without careful consideration of the potential value to be obtained through utilizing the different types of designs. The advantages to mixed methods research suggested by Johnson and Onwuegbuzie (2004) will only be obtained through careful design. As suggested above, the use of mixed methods requires clear understanding of the philosophies underlying the method and how the method is employed as well as the strengths and limitations of the method. For example, interviewing is a method, but phenomenological interviewing is a methodology, a method embedded in a philosophical approach. Additionally interviewing has strengths and weakness in comparison to other methods such as surveys. In order to capitalize on the strengths of mixed methods designs, evaluators must recognize and be explicit about their selection and use of evaluation methods. Otherwise, the use of mixed methods could degenerate into using whatever approach is easiest or most convenient.

This chapter will explore the different models of mixing methods using various NSF evaluations as examples. The philosophies underlying different methodologies and the strengths and limitations of different methods will be highlighted. Suggestions to promote the thoughtful use of mixed methods will be provided and future research on the use of mixed methods in STEM evaluation will be proposed.

References:


One of the key issues in multisite STEM evaluation is the relationship between evaluation use and results. The accountability movement and the focus on scientific evaluation has placed emphasis on determining measurable goals for program activities, and measuring performance against goals. Equally important in this whole movement; however, is evaluation use and its relationship to program results. In this manuscript we use a program-level evaluation to assess the characteristics of smaller-scale, project evaluations and their use for project improvement and accountability purposes. Many federal and private foundations conduct large programs that fund multiple projects. These programs are often evaluated for both program level development and accountability purposes. Routinely, these evaluations place their evaluation lenses on learning about (describing) the program and its accomplishments. While the focus of these evaluations is typically on program objectives and activities, often they gather data about evaluation practices.

Program level evaluations that gather information about project-level evaluations provide particularly useful entrées to studying evaluations themselves and their contributions to projects. A program has at least two important characteristics that facilitate this focus, (a) the funded projects have common program-based goals and objectives and (b) expectations for project-level evaluations derive from a common set of requirements or guidelines serving all projects. For those conducting programs such studies of evaluation practices can help them better understand how evaluations serve their programs and ways in which evaluations can be improved. In a more general sense such studies contribute to our knowledge of evaluation practices and help us to better understand practical aspects of evaluations and their use and can lead to improved evaluation theory and instruction (Mark, 2004).

Here we used our Evaluation of the NSF Advanced Technology Education (ATE) program to shed light on the use of evaluation at the project level. We focused on two points of interest. First, we addressed questions related to the nature and extent of evaluations conducted at the local level. We sought to better understand what constitutes evaluation at the project level. Second, we sought to learn whether and to what extent such evaluations serve project productivity (i.e., are useful to projects). As such the study fits within two inquiry modes, description and causal analysis, described by Mark (2004) as ways of building a better evidence base for evaluation theory.

To conduct this description and causal analysis, we conducted a secondary analysis of survey data from the ATE program. As such the questions we could ask and answer were constrained
by the data already obtained through survey efforts. We could not and did not generate hypotheses first and then gather data to address those questions. Rather, we reviewed the survey questions answered by project staff to determine what of the collected information addressed evaluation practices and then studied relationships between those variables and the output and outcome information projects provided. This type of analysis is less than optimum for fully assessing the characteristics of local project evaluations. But, in the case we investigated, the program evaluation did collect substantial information regarding local project evaluations. As such, it provides a good and interesting mechanism for studying project evaluations and their viability.

The ATE program funds approximately 250 projects and centers with a current total annual budget of just less than $40,000,000 per year. The 2004 survey data used was the fifth annual survey of projects conducted as part of an external evaluation of the ATE program being conducted by The Evaluation Center at Western Michigan University. We surveyed the population of currently funded centers and projects that have received funding for at least one year. These data were used to provide an annual status report, a description of the current work and productivity of the ATE program through aggregation of information across the individual projects and centers.

The Web-based survey sample was intended as a census of the target population. The response rate of 154 represents a response rate of 97%. However, not all respondents completed all items; this occurred primarily because the survey was separated into 7 sections with 3 parts required for all respondents and 4 dedicated to projects conducting specific types of activities. The four areas of specific activities included: materials development, professional development, program improvement, and articulation agreements. A number of questions in both the overall and activity specific sections pertained directly to evaluation and assessment matters.

The premises of the ATE program and our evaluation were that (a) evaluation is important for program accountability purposes and (b) evaluation also serves projects’ objectives, processes, and outcomes.

Two sets of indicators were developed from the 2004 survey (a) evaluation indicators and (b) indicators of project performance. Among the evaluation indicators, there were three distinct types: (1) general evaluation indicators, and evaluation indicators specific to two types of activity (2) materials development and (3) professional development. Five types of performance indicators were generated including: (1) collaboration, (2) materials development, (3) professional development, (4) program improvement, and (5) articulation agreements. As those labels suggest, the performance indicators are consistent with program expectations and project activities.

We began this investigation by determining the proportion of projects that report conducting evaluations. Then we followed that initial response to gain some indication of what comprises evaluation efforts for these projects. To do this, we identified and organized data that we considered to be evaluative in nature (e.g., whether the project collected end-of-program reaction data for professional development workshops). From those data we constructed evaluation effort
profiles across projects (e.g., on average 4.9% of the annual project budget is spent on evaluation).

Finally, we looked at our evaluation characteristics in conjunction with the multiple project performance indicators. Here we were looking for ways in which evaluation might be related to enhanced or decremented project results. We identified outcome variables in the five categories described above. For each outcome variable we used multiple regression to test the extent to which evaluation practice and perceptions of evaluation usefulness predicted the outcome. For all five categories the general evaluation indicators served as predictors. For materials development we also used the nine materials development evaluation indicators as predictors. Similarly, for professional development we additionally used the three professional development evaluation characteristics as predictors. For each category significant R values were obtained. These relationships between evaluation efforts and project performance provide insights to both the value of evaluations and ways that evaluations can be improved to better serve their projects and the program as a whole.

References:

Overview
This article will review program evaluations that The National Science Foundation’s (NSF) Division of Research, Evaluation and Communications have carried out in recent years. Evaluations have reflected the field-driven nature of NSF’s programs, guided primarily by peer review. There are also other factors that must be taken in consideration when examining evaluations that have been conducted, such as the nature of the program and the program status. These have created limitations for evaluations that must be accepted or other program rationales must be developed.

Recent Status and Rationale for Program Evaluation
NSF has been conducting some program evaluations since the mid-1970s. Most were conducted by the Education Directorate of NSF rather than the research directorates. The budget for program evaluation has grown substantially since that time; however, the pattern of growth has been uneven. After building through the 1970s, the Education Directorate lost most of its funding under the Reagan Administration and program evaluations ceased. When education was rebuilt later in the 1980s, program evaluation returned and its budget increased steadily through the 1990’s, reaching its high of $12.5 billion in 2003, when it began to decrease. It is now uncertain what the evaluation budget will be in the coming years, or even if there will be one.

NSF has been primarily a granting agency, establishing broad research programs that fund specific proposals primarily on the basis of peer review. The same rationale is used with educational programs, although it may be less compelling in education where there is less likely to be agreement in the professional community on desirable next steps to support. The result is programs (educational as well as research) with very broad goals and objectives. To then try to specifically identify for evaluation what those programs are supposed to accomplish is a difficult, if not impossible task.

However, it is important to recognize that this approach reflects a fundamental belief that research and education proposals should be funded in a manner that provides the greatest flexibility to proposers and the field in determining priorities. To require specific objectives or
outcomes for the programs would be seen as the evaluation tail wagging the NSF dog. Evaluation must fit within this belief system of the scientific community to have any hope of existing at NSF.

There are also other factors that must be taken into consideration when examining NSF evaluations, such as: What is the nature of the program? Although NSF education programs are similar in process, they represent a wide range of purposes from fellowships that don’t specify any specific educational program to education interventions with limited scope to comprehensive educational programs that are difficult to separate into components.

What is the program status? For programs in their infancy, concerns are likely to be in implementation while outcomes can be emphasized for more established programs. What are the priorities for evaluation? At one time, NSF was required to evaluate all education programs on regular cycle. Later, an emphasis was placed on evaluating new programs when program adjustments could more readily be made. There were also attempts to accommodate a range of methodologies. And there were always specific program concerns or emphases expressed by various levels in the NSF hierarchy or beyond.

This paper will include analyses of the evaluations conducted through the Research, Evaluation and Communication Division of NSF that have either been recently completed or that are still operating. It will be possible to explore the rationale for each evaluation and to discuss the findings for a number of them. Finally, this article will explore alternative approaches that NSF might consider for future evaluations of STEM programs.
The provisions in the No Child Left Behind Act (NCLBA) call for increased emphasis on scientifically based evidence (Mertens & McLaughlin, 2004). The NCLB sets the use of standardized tests and randomized designs using the scientific method as the desired approach to demonstrate a program’s effectiveness (U. S. Department of Education, 2004) (USDOE). In response to USDOE’s elevation of the scientific method and randomly controlled trials, the American Evaluation Association (AEA) (2003) stated, “While we agree with the intent of ensuring that federally sponsored programs be "evaluated using scientifically based research . . . to determine the effectiveness of a project intervention," we do not agree that "evaluation methods using an experimental design are best for determining project effectiveness."

AEA takes the position that there is not one right way to evaluate the effectiveness of a program. AEA is joined by other organizations such as the National Education Association (NEA) in providing commentary on NCLBA. The NEA communicated with then US Secretary of Education, Rod Paige, cautioning that we need to use a multiplicity of approaches, including the scientific method, when feasible, to demonstrate effectiveness of programs. The position specifically states, “(1) the evaluation approach used be appropriate for the problem or question the program itself seeks to address; (2) that the evaluation definition and set of priorities used are not so narrow that they effectively preclude the funding of worthwhile programs; and (3) that the Department continue to recognize the importance of third party, independent evaluators” (Moody, 2003, p. 1).

The Transformative Prism
The world of evaluation can be seen as trying to understand the reality of educational or social programs as through a prism. The prism refracts the differences of experiences into an ever-changing pattern of different lights, while we seek ways to understand the use of culturally appropriate, multiple methods in understanding the pattern of diverging and converging results of the evaluation. The transformative paradigm provides a useful theoretical umbrella to explore the philosophical assumptions and guide methodological choices for the approaches to evaluation that have been labeled inclusive, human rights based, democratic, empowerment or responsive. The transformative paradigm extends the thinking of democracy and responsiveness by consciously including the identification of important dimensions of diversity in evaluation work and their accompanying relation to discrimination and oppression in the world (Mertens,
2005). It prompts the evaluator to ask such questions as: What is hidden in the mandate of scientifically-based research and use of "reliable" and "valid" standardized tests when applied to populations that are extremely diverse and not found in large groups that can be ethically or logistically randomly assigned to conditions? What is the evaluator's role in uncovering that which has not been stated explicitly and the danger that lurks in applying the conceptualization of scientifically based research without consideration of important dimensions of diversity?

There are three sets of philosophical assumptions that are most relevant to defining a paradigm in an evaluation context:

- reality (ontology)
- relation between knower and would-be-known (epistemology)
- the appropriate approach to systematic inquiry (methodology)

Ontologically speaking, how do we know that something is real? I don’t mean a table or a computer that I can touch. I mean the realities that we know at a conceptual level, for example when is access real? When is an environment least restrictive? When is literacy real? In the ontological sense, you have an assumption about what is real when you decide what type of evidence you will accept that someone is indeed literate or any other conceptual characteristic.

Epistemologically, you ask yourself: If I am to really know if something is real, how do I need to relate to the people from whom I am collecting data? So the knower is the evaluator and the would-be-known is the subject or participant in the study. Should I be close to the participants so I can really understand their experiences or should I maintain distance between myself and the subjects so I can be neutral? This question of course raises the definition of objectivity as it is operationalized in an evaluation context.

Methodologically, I have choices to make that go beyond quantitative or qualitative or mixed methods, to how do I collect the data about the reality of a thing in a way that I can feel confident that I have indeed captured that reality?

Cultural Competency  Inherent within the philosophical assumptions of the transformative paradigm is an underlying principle that relates to the concept of cultural competency (Mertens, 2005). Multiple definitions of cultural competency are found in the scholarly literature. For example:

➢ Cultural competence refers to an ability to provide services that are perceived as legitimate for problems experienced by culturally diverse populations. This definition denotes the ability to transform knowledge and cultural awareness into interventions that support and sustain healthy participant-system functioning within the appropriate cultural context. (Guzman, 2003, p. 171)

➢ As an agent of prosocial change, the culturally competent psychologist carries the responsibility of combating the damaging effects of racism, prejudice, bias, and oppression in all their forms, including all of the methods we use to understand the populations we serve…A consistent theme…relates to the interpretation and dissemination of research
findings that are meaningful and relevant to each of the four populations (Author added: Native American, Latino, African American, and Asian American) and that reflect an inherent understanding of the racial, cultural, and sociopolitical context within which they exist. (APA, 2000, p. 1).

- A culturally competent evaluator is one who is actively in the process of becoming aware of his or her own assumptions about human behavior, values, biases, preconceived notions, personal limitations, and so forth. Second, a culturally competent evaluator is one who actively attempts to understand the worldview of the culturally different participant group. In other words, what are the values and assumptions about human behavior, biases, etc. of the group(s) being evaluated? Third, a culturally competent evaluator is one who is in the process of actively developing and practicing appropriate, relevant, and sensitive evaluation strategies and skills in working with culturally different groups. These three goals make it clear that cultural competence is an active, developmental and ongoing process and that it is aspirational rather than achieved. (Sue & Sue, 2003)

Symonette (2004) reminds us that cultural competency is not a state of being, rather it is a journey. “…Evaluators need enhanced understandings of related systemic processes of asymmetric power relations and privilege, not simply awareness and knowledge of difference and diversity….How and to what extent is sociocultural diversity associated with patterned differences in access, resource opportunities, and life chances?” (p. 108)

Many evaluation leaders are careful to point out that cultural competence cannot be determined by a simple checklist, but rather it is an attribute that develops over time. The root of cultural competency in evaluation is a genuine respect for communities being studied and openness to seek depth in understanding different cultural contexts, practices, and paradigms of thinking. This includes being creative and flexible to capture different cultural contexts, and a heightened awareness of power differentials that exist in an evaluation context. Important skills include: ability to build rapport across difference, gain the trust of community members, and self-reflect and recognize one's own biases. (Edno, Joh, & Yu, 2003).

Frierson, Hood and Hughes (2002) outline specific considerations at each stage of an NSF evaluation that are designed to improve its cultural responsiveness. These points occur in the preparation stage, the engagement of stakeholders, the identification of purpose and intent, framing the right questions, design of the evaluation, selection and adaptation of instrumentation, the process of data collection, the analysis of data, and the dissemination and utilization of results. This chapter will examine these stages of the evaluation process in terms of the strategies for improving the cultural responsiveness of STEM evaluations within the context of STEM and other NSF funded initiatives.

References


Social Justice Issues in STEM Educational Evaluation\(^1\) – Jennifer C. Greene, Lizanne DeStafano, Holli Burgon, Jori Hall

Whether motivated by a desire to improve the human condition or to remain at the forefront of other industrialized nations, there is a strong press among policymakers, industry leaders and educators to improve the quality of math, science and technology education at all levels and to increase the numbers of students who are interested in STEM fields, particularly among groups who have not traditionally chosen STEM careers, notably, women, ethnic and racial minorities and persons from lower socioeconomic groups. The persistent lack of diversity in science classrooms and laboratories is not only socially unjust, it also compromises the vitality and creativity of STEM endeavors. Inventions, breakthroughs, and significant progress in science-related understandings and applications are much less likely to happen under conditions of homogeneity of thought and perspective.

This chapter reports on an approach to STEM education evaluation that seeks to contribute to our collective understanding of STEM practices that show promise of meaningfully increasing active participation in STEM fields by diverse learners and that offer bold and new ways of thinking about science, education and diversity. This evaluation approach is anchored in two principal commitments – to evaluation as an educative practice and to evaluation as a forum for engaging with critical values. The approach also features three supporting perspectives or tools, through which the principal commitments are importantly enacted – contextualization, inclusion, and program theory. Each of these commitments and perspectives will be elaborated on in the final manuscript.

Evaluation is fundamentally educative. This evaluation framework seeks to engender deep understanding among local program staff and consumers, as well as the field at large, regarding the salient components of a STEM program, how they interact with each other and with contextual factors, and the resulting nature and level of program outcomes. This approach aims to describe the pedagogical quality of the learning experiences of students and teachers in the program being evaluated, to understand the STEM disciplinary thinking and perspectives offered

\(^1\) This work is supported by NSF EREC grant #0335621. We thank the members of our evaluation advisory board Susan Cohen, Frank Davis, Stafford Hood, Frances Lawrenz, Tom Schwandt, and Bob Stake for their thoughtful contributions to our work.
by the program, and to probe the ways in which the program at hand serves the interests of all learners in that context, especially learners from groups underrepresented in STEM fields. In this way, this approach offers an opportunity to learn about the particular contextual character and contours of meaningful, high quality, effective STEM education programs at the intersection of science, pedagogy and equity.

**Evaluation is value-engaged.** Increasing the magnitude and quality of the STEM pipeline is dependent, in part, on the ability of educational institutions in our country to reposition themselves in order to effectively address the needs of the underserved. As noted, judgments of the quality of STEM programs must consider explicitly, in addition to the quality of science content and pedagogy, the extent to which a program has been successful at redressing past inequities and increasing the participation of underrepresented groups in STEM fields. That is, in this evaluation approach, STEM education programs are importantly judged by how well they engage, include, respect, and encourage students from under-represented groups, along with the perspectives, life experiences, practical wisdom, and understandings about science, technology, engineering, and mathematics that under-represented students bring with them to the teaching-learning context.

This stance demands an active engagement with values and redefines the role of the evaluator as someone who continually and knowingly reintroduces values into the evaluation process. This can occur through a variety of means, many drawn from responsive and democratic evaluation traditions, such as inclusion of various stakeholders in the evaluation process, representation and valuing of multiple and diverse perspectives of quality, dialogue among stakeholders to promote understanding, and representation of both minority and majority interests in the evaluation.

**Enacting educative, value-engaged evaluation.** Three evaluative perspectives hold promise for enabling evaluation that is educative and value-engaged.

♦ **Contextualization.** Evaluation in this approach assesses how well a STEM educational program offers meaningful learning opportunities and meets important educational needs of teachers and learners in their particular context. Moreover, this evaluation framework also examines the extent to which program features meet learners’ needs as a way of interpreting and improving learner outcomes. That is, most approaches to educational evaluation assess how well students perform in the program being evaluated, and this is clearly important. In addition, in the value-engaged approach, evaluations assess how well the program “performs” in a particular context, or how well it fits the people and their expectations, the culture, the daily rhythms and routines, the stresses and tensions of the particular context at hand – in terms of the program’s design, implementation, and impact (Greene, 2004; Greene, Millett, and Hopson, 2004; Kushner, 2000).

♦ **Inclusion.** In educative, value-engaged evaluation, the interests and perspectives of all legitimate stakeholders are included, with particular attention to the interests and perspectives of those traditionally under-served by our educational systems and under-represented in STEM policies, programs, and practices, including racial and ethnic minorities, women and girls, low-income learners, immigrants, and learners with disabilities. The interests of the majority are not excluded in this evaluation approach; rather the interests of the minorities
are specifically included (House and Howe, 1999). By actively seeking to represent the interests and perspectives of all legitimate stakeholders, particularly those traditionally underserved, evaluation can reveal the multiple and diverse perspectives, experiences, and impacts associated with a single program and increase the depth and breadth of understanding of STEM program operation and effectiveness.

♦ **Program Theory.** Finally, educative, value-engaged evaluation is oriented around program theories. The questions and concerns addressed in this approach to evaluation are importantly connected to and framed by the logic of the program as designed or intended and as experienced by participants. Program theories in this approach, that is, are multiple and contextual. They provide handles or levers for engaging with various understandings of STEM content and pedagogy, with ideas about how to make STEM contexts more welcoming of diversity, and especially with the assumptions and associated value stances that underlie these various understandings. Explicit program theories further enable an assessment of the extent to which values of social justice, inclusion and diversity are manifest in the theories espoused by different stakeholders or the likelihood that the various theories converge to produce a set of commonly valued outcomes.

The chapter will present these conceptual components of the educative, value-engaged evaluation approach and then illustrate them with an extended example.

**References:**


The purpose of this manuscript is to describe a new model of evaluation capacity building in STEM education. The new model we present will help to further define the concept of evaluation capacity building, and will help to encourage new ways of thinking about evaluation capacity building. The concept of evaluation capacity building is difficult to define. Compton, Braizerman, and Stockdill (2002) recently provided a framework of the evaluation capacity building concept and a working definition. They defined evaluation capacity building as “the intentional work to continuously create and sustain overall organizational processes that make quality evaluation and its use routine (p. 109).” They also described eleven different elements of evaluation capacity building in an attempt to provide a comprehensive overview of the complex nature of evaluation capacity building. However, as Compton, Braizerman, and Stockdill point out, much work still remains to further the evaluation capacity building concept. One way to move the concept of evaluation capacity building forward is by providing models of and testing out those models in real-world settings. The intent of this manuscript is to present such a model. We were recently funded by The National Science Foundation to develop the Collaborative Evaluation Communities in Urban Schools project. This project uses an immersion approach to develop the evaluation capacity of urban schools, and to develop the evaluation capacity of STEM graduate students. The project provides a unique model of how to immerse K12 teachers and graduate students in the evaluation process in an attempt to develop the evaluation capacity of both groups.

---

1 This work is supported by a grant from the National Science Foundation, EREC #0438069
Building the evaluation capacity of K12 schools is clearly an important goal for the field of evaluation, especially in the current educational environment that is dominated by issues of accountability and high stakes testing. King (2002) provided a case study of her work on evaluation capacity building in a large suburban school district. The case illuminates the challenges confronting large school districts as they attempt to increase their capacity to conduct and use program evaluation. K12 schools are facing an onslaught of tests designed to judge whether or not they are making adequate yearly progress (AYP) in increasing student achievement. Currently mathematics and reading test scores are the main measures of AYP, but starting in 2007 science test scores will also be required as part of the No Child Left Behind Act. The tests scores are made public for all to see, and schools that do not make adequate yearly progress are singled out and risk being placed on a ‘needs improvement’ list or even worse subject to external intervention. The focus on student test scores has forced schools to become more data-driven as they attempt to analyze test scores and make decisions about how they can improve scores next year. To some extent, schools have become overwhelmed with data. State assessments, district assessments, schools assessment and classroom assessments have all led to more data than schools can reasonably manage. One could argue schools especially need to develop evaluation capacity to help manage the situation. The schools not only need to develop better evaluation infrastructure to handle the multitude of data, but they also need to help administrators and teachers alike develop evaluation knowledge and skills, and increase their ability to conduct program evaluation.

Collaborative Evaluation Communities in Urban Schools (CEC): In this manuscript we will provide a description of the Collaborative Evaluation Communities in Urban Schools project (CEC) and how the project builds evaluation capacity of participants. The CEC model is based upon an immersion approach to evaluation capacity building. In the CEC model both K12 teachers and graduate students are immersed in the evaluation process as a means of developing their evaluation capacity. Through this model we seek to build a culture of evaluation in the schools where all the participants are engaged as a community in evaluation. The process of engaging in a collaborative evaluation community helps participants improve their ability to collect and analyze data, examine student assessment results, evaluate school programs, analyze instructional practices in science and mathematics classrooms, and ultimately build infrastructure for data collection and analysis. The collaborative evaluation communities also serve as training grounds for STEM graduate students. The field of science and mathematics education is in need of more science and mathematics educators with evaluation expertise. The CEC project helps to develop new evaluators by immersing graduate students with science or mathematics expertise in collaborative evaluation communities, and by providing coursework and professional learning opportunities. In addition, the collaborative evaluation communities provide professional growth activities for elementary and middle school science and mathematics teachers. By engaging teachers in the evaluation process we can assist partner schools in the development of an evaluation culture designed to support teachers in the continual examination of programs with the ultimate intent of improving the educational opportunities for all students.

The key theoretical concept behind the CEC model is that immersing teachers in the evaluation process can help build evaluation capacity and help bridge the gap between district evaluation efforts and the teaching and learning of science and mathematics. Collaboration has been shown
to be a powerful tool for influencing individual beliefs and practices (Bissex, 1994; Cochran-Smith & Lytle, 1993; Kalnin, 2000). Collaborative efforts that engage individuals throughout a district in a shared evaluative process of data-gathering and analysis can lead to sustained improvements in science and mathematics teaching and learning (Huffman & Kalnin, 2003). Collaborative efforts support participants in building contextualized knowledge of their students, their science and mathematics programs, and the school community. Cochran-Smith and Lytle (1999) describe such collaboration as essential to developing knowledge. They contend, “Knowledge emerges from the conjoined understandings of teachers and others committed to long-term highly systematic observation and documentation of learners and their sense making” (p. 275). Sarason (1996) urges schools and universities to create opportunities for collaboration that move across current hierarchies. He concludes, “Teachers cannot create and sustain contexts of productive learning for students if those contexts do not exist for teachers” (p. 253-254). Structures that engage schools, teachers, graduate students and professors in joint investigations create opportunities to share expertise and build interdependence in understanding evaluation issues within a particular context (Elmore & Burney, 1999; Fullan, 1993; Little, 1999; Sarason & Lorentz, 1998).

Conclusion: In the face of increasing emphasis on accountability and requirements to continually improve student achievement in mathematics and science, educational practices in K12 schools are placed at the forefront of STEM evaluation. The No Child Left Behind Act mandates that K12 schools demonstrate AYP in increasing student achievement. As a result, administrators and teachers in schools must develop the capacity to continually evaluate educational programs. One method for accomplishing this goal is for STEM evaluators to initiate evaluation capacity building efforts in K12 settings. The complex nature of evaluation capacity building presents challenges for schools that attempt to increase their capacity to conduct and use program evaluation (King, 2002). The CEC model described in this manuscript will draw on the framework of the evaluation capacity building concept presented by Braizerman, Compton, and Stockdill (2002). This project moves the concept of evaluation capacity building forward through the application of a collaborative model for evaluation capacity building in real school contexts. The CEC model places evaluators, graduate students, and school personnel closer to the contexts of the educational settings and supports the continuous refinement of evaluation efforts based on student learning of science and mathematics. This model supports the development of evaluation capacity through collaborative evaluation communities that can move beyond the current organizational hierarchies in schools that have hampered change efforts in the past, and move schools towards new ways of evaluation capacity building.

References


The final chapter in the volume will provide a critical review of the material presented in each of the chapters in relation to the ideas presented in the introductory chapter. It will discuss the importance of the different issues raised in the various chapters to the conduct of STEM evaluation and to evaluation in general. The ideas will be discussed in the context of “scientific” evaluation and what constitutes rigor as well as how the ideas contribute to notions of causality. The chapter will include a discussion of each of the chapters separately and then an overarching review of the main issues that emerged and their relevance. It will also make recommendations for future theoretical and empirical work in these areas.