USING MULTI-SITE CORE EVALUATION TO PROVIDE “SCIENTIFIC” EVIDENCE

Frances Lawrenz
University of Minnesota
Minneapolis, Minnesota

Douglas Huffman
University of Kansas
Lawrence, Kansas

Abstract: Funders of educational and other social service programs are requiring more experimental and performance-oriented designs in evaluations of program effectiveness. Concomitantly, funders are exhibiting less interest in evaluations that serve other purposes, such as implementation fidelity. However, in order to fully understand the effectiveness of most complex social and educational programs, an evaluation must provide diverse information. This article uses the Core Evaluation of the Collaboratives for Excellence in Teacher Preparation Program as a case example of how evaluations might meet the requirements of objective scientific evaluation while at the same time valuing and incorporating other evaluation purposes. The successes and limitations of the case example in achieving this blending are discussed.


Corresponding author: Frances Lawrenz, Chair, Educational Psychology, Wallace Professor of Teaching and Learning, University of Minnesota, 178 Pillsbury Drive SE, Minneapolis, MN, USA 55455; <lawrenz@tc.umn.edu>
Entities commissioning social service evaluations, such as the Canadian and United States governments and other public and private funders, are becoming more and more interested in “scientific” evaluation of program effectiveness. In fact, the U.S. Elementary and Secondary Education Act (ESEA) as reauthorized by the No Child Left Behind Act of 2001 (NCLB) uses the term “scientifically based” more than 100 times in the context of evaluating programs. 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 their annual budget submission. Similarly, the Canadian Budget Transparency and Accountability Act mandates performance-based public accountability (McDavid, 2001). Against this background, commitment to accountability has expanded. Accountability may mean different things, such as development of standards, documentation of program processes, improvement of program offerings, indication of programs doing what they said they would do, or achievement of specific outcomes. In the present climate, however, the meaning of accountability is narrowing to achievement of specific outcomes.

In 2002, the U.S. Office of Educational Research and Improvement was reconstituted into the Institute of Education Sciences, reflecting the intent of the U.S. government to advance the field of education by requiring more rigorous educational research and supporting implementation of evidence-based programs. In the fall of 2003 the U.S. Department of Education issued a notice of proposed priority on “scientifically based evaluation methods” (U.S. Department of Education, 2003). The Secretary of Education proposed a priority for projects employing an evaluation plan that is based on rigorous, scientifically based research methods. The document defines “scientific” methods in a priority list.

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. (U.S. Department of Education, 2003, p. 2)
Next on the priority list of evaluation methods are quasi-experimental designs with matched comparison conditions, which may be used when random assignment is not possible. 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 causal relationships.

Despite the U.S. government’s assertion of the value of experimental design-based research and evaluation, a recent report from the U.S. National Research Council’s Committee on Scientific Principles for Education Research, Scientific Research in Education (NRC, 2002) supported a broader notion of research. The debate about what would constitute adequate measures of program effectiveness is supported by contrasting paradigms of research and evaluation. Different approaches to evaluation are grounded in different philosophies. House (1983) has categorized these differing philosophies along two continua: the objectivist-subjectivist epistemologies and the utilitarian-pluralist values. Objectivism requires evidence that is reproducible and verifiable. It is derived largely from empiricism and related to logical positivism. Subjectivism is based in experience and related to phenomenologist epistemology. The objectivists rely on reproducible facts while the subjectivists depend upon accumulated experience. In the second continuum, utilitarians assess overall impact while pluralists assess the impact on each individual. In other words, the greatest good for utilitarians is that which will benefit the most people while pluralism requires attention to each individual’s benefit.

The assertion that “scientific” evaluation is the best way to determine program effectiveness marginalizes other philosophical stances. As Maxwell (2004) points out, 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.

Feuer, Towne, and Shavelson (2002) and others (Eisenhart & Towne, 2003; Jacob & White, 2002) argue that emphasis should be placed on understanding and appreciating multiple perspectives rather than advocating particular methods. Stufflebeam (2001) categorized over 20 different evaluation approaches, assessed their strengths and weaknesses, and considered whether, when, and how the methods are best applied. Despite this categorization, there is little empiri-
cal evidence supporting the superiority of one method of evaluation over another. 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). Furthermore, using “scientific” methods may not produce more objective results. Philosophers of science have written at length about the social, political, and subjective nature of science and point out that scientists tend to gather data to support their ideas and beliefs, rather than objectively collecting, analyzing, and attending to data (Jahn & Dunne, 1997; Kuhn, 1977; Malhotra, 1994; Popper, 1972; Scheffler, 1967).

The nature of social service innovations poses important difficulties in their evaluation. Social service innovations are generally complex programs including many variables and having diverse goals and outcomes (Berliner, 2002). The outcomes are often unique, value-laden, and with varying end points. If a specific outcome is agreed upon in advance as the most valuable, evaluation is more straightforward; however, this type of agreement rarely occurs. For example, even in a medical trial the goal is seldom as simple as “remaining alive”; the goal is “wellness” and a myriad of variables might be valued outcomes of such a program. Additionally, what wellness means to different people or groups is likely to differ. The very nature of social service programs is more subjective and pluralistic than the objectivist, utilitarian nature of experimental design.

It is generally conceded that in order to conduct rigorous evaluations of the effectiveness of social service programs, an evaluator needs the cooperation and participation of the various stakeholders (Fitzpatrick, Sanders, & Worthen, 2003). This participation allows the collected data to have more fidelity to the actual innovation, be more complete because the participants are willing to provide it, be more valid because participants understand what data are necessary, be more fully interpreted because more people are involved, and be used more. Additional use results in greater dissemination of the ideas, more implementation of the successful parts of the innovation, and better programs. Cousins and Whitmore (1998) describe practical participatory evaluation as one that is common in the United States and Canada and “has as its central function the fostering of evaluation use with the implicit assumption that evaluation is geared toward program, policy, or organizational decision making” (p. 6).
Given the recent emphasis on scientific approaches, it is essential that the field find ways to integrate a scientific approach with other methods to assess social service program effectiveness. The purpose of this article is to describe and critique an evaluation that merged scientific and participation principles, the Collaboratives for Excellence in Teacher Preparation Core Evaluation project. Through consideration of the successes and limitations of this case, ideas on how best to accomplish the merger can be made.

THE CETP PROGRAM

The Collaboratives for Excellence in Teacher Preparation (CETP) is a program funded by the U.S. National Science Foundation (NSF). According to the NSF, the CETP program promotes comprehensive change in the undergraduate education of future teachers by supporting cooperative, multyear efforts to substantially increase the quality and number of teachers well-prepared in science and mathematics, especially members of traditionally underrepresented groups. Nineteen system-wide CETPs were funded from 1993 through 2000, when the program stopped receiving new proposals. Projects were funded for five years with the option for an additional three. These were large projects of several millions of dollars each.

Although each CETP was unique, they all pursued the overriding goal of the program stated above. There were many characteristics common to all CETPs. All CETPs engaged faculty from the sciences, mathematics, and education. All included several institutions of higher education (including community colleges) in a particular geographic area (e.g., a state). All included mechanisms for improving undergraduate education in the sciences and mathematics. All had some sort of relationship with schools within the geographic area. All offered scholarships to students from underrepresented groups. The ranges within these commonalities, however, were broad (e.g., from 3 institutions to 13 or more), and all had unique elements (e.g., liaisons with Tribal Colleges or industrial/science internship placements).

THE CETP CORE EVALUATION

The CETP Core Evaluation was a separate project from the CETPs funded in 1999 after the CETPs were in operation. The evaluation question directed at the CETP Core Evaluation project by NSF was:
What evidence exists that the changes instituted as part of the Collaboratives have indeed resulted in a substantial increase in the number of students who know more, are more competent at teaching mathematics and the sciences using the mathematics and science standards as a guide, and employing the new technologies available?

The Core Evaluation project instituted a process to develop an evaluation that included “scientific” evaluation ideas. It was necessary for the Core Evaluation project to use a participatory approach because collaboration was not required by NSF; it was only encouraged. The project, therefore, instituted an approach in which all of the CETPs could feel ownership and be willing to contribute. The participatory approach used in CETP is slightly different from the type of participatory evaluation described by Cousins and Whitmore (1998), which has evaluators working in a partnership with others who are not so trained, such as members of the program community. In the CETPs, the local evaluators served as internal evaluators for their own projects. They had a variety of roles and backgrounds, including co-principal investigators, coordinators, faculty members, and graduate students. The approach was participatory in the sense that all would gain from the evaluation and all had input into the design, analysis, and so on. It could perhaps be considered more of a collaborative evaluation than a strictly participatory one.

First, the features of the program to be evaluated were determined. This was accomplished through extensive involvement of the CETP sites. The plan was to develop consensus on the evaluation questions, evaluation design, sources of data, data collection devices, and report development. It was hoped that development and implementation of the core evaluation would allow the local sites to concentrate their evaluation efforts on the unique elements of their projects while the Core Evaluation project concentrated on the common elements that could be addressed using quasi-experimental designs. A meeting with the CETP evaluators was held soon after the Core Evaluation project was funded, with the Core paying the CETP evaluators for consulting time and travel expenses. The first meeting concentrated on developing evaluation questions that all CETPs felt addressed the goals of their individual projects. This also included the specification of sources of information and a sampling plan. A quasi-experimental design was developed, focusing on two settings: K–12 schools and institutions of higher education. These settings matched the goals of the CETP program, which were to improve the
preparation of teachers through the improvement of college-level coursework and to improve the education of K–12 students through the provision of better prepared teachers. At the meeting evaluation questions to be addressed by the Core evaluation were agreed upon. These were:

Ia. How supportive of science, technology, engineering, and mathematics (STEM) reform education policies and procedures are the participating CETP institutions?
Ib. How successful have the CETPs been at course reform?
Ic. What impact have CETPs had on the system or structure of the teacher education systems at the participating institutions?

IIa. How well do CETP teachers demonstrate the knowledge and skill espoused by the standards?
IIb. How do CETP teachers and the classrooms they create differ from non-CETP teachers and the classrooms they create?
IIc. What outcomes have the participating higher education institutions, their faculty, or the CETP contributed to the K–12 schools?

IIIa. Are students learning what is expected in the STEM standards?
IIIb. Are there differences in student outcomes for CETP and non-CETP teachers?

The “scientific” portions of the overall evaluation design in the institutions of higher education were either pre-post (where the faculty were asked about how they taught before and after the CETP) or comparison-to-standards assessment (where the perceptions were compared to an ideal) using a random selection of faculty to determine spread of ideas across institutions, while the design at K–12 was comparisons of matched groups. As can be ascertained from the evaluation questions, within each setting the “scientific” data collection was supplemented by descriptive components and triangulated by obtaining similar information from different sources. Details of the design within each setting follow.

Higher Education Setting

Data in the higher education settings were to be gathered from a subset of science, technology, engineering, and mathematics (STEM) faculty, all STEM education faculty, all STEM and STEM educa-
tion deans and/or department chairs, and the principal investigator (PI) for the project. A sample of three STEM faculty members was to be obtained per department per institution. So each two- and four-year institution in each collaborative was to have three members from the physics department, three from the chemistry department, and so on, to be randomly selected from the regular teaching faculty. This allowed the estimation of the amount of spread of the innovative teaching ideas throughout the faculties at the participating institutions. All faculty were to be asked to reflect on how they had been involved in education and how they were involved now. The PIs, deans, and department chairs were to be asked about present status and changes that had occurred.

In addition, 15 courses, which could be used by students to meet the requirements in the teacher preparation program(s), were selected to be observed. These were to be spread out by institution type in the collaborative (e.g., five per large research institution, five per four-year college, and five per two-year college; or seven per four-year college and eight per two-year college), to make a total of 15. The observations were to be of only one class session and for approximately one hour total, although the observer could go in and out of long classes. The faculty members teaching these courses and the students in them were also to be surveyed. Because class mean data were the goal, student questionnaire items were to be divided within two forms so only half the students in each course completed each item. This shortened the amount of time necessary for students to fill out the survey. Artifacts of a class activity and a class assessment from each of the 15 courses were also to be gathered.

K–12 Setting

Graduates of the CETP program were to be followed and contacted after completing their first year of teaching, obtaining a random sample of 30 CETP-trained teachers. A matched teacher (same grade level and/or content) was to be selected from that same school or a matched school. The teachers (CETP and non-CETP), the principal, and the students of the teachers were to be sent surveys to complete. Teachers were also to be asked to send in two artifacts: a student activity and a student assessment. A subset of 12–16 of the teachers (those in the same or best-matched schools) was to be selected for observation. The sample was to contain six to eight CETP teachers and six to eight non-CETP teachers. Teachers in their second or third year were to be selected to allow their teaching style to stabilize.
Instruments

Once the evaluation design had been determined, instruments were developed. Instruments from all of the CETP's had been gathered before the first meeting. These were categorized and sent out to staff at the various CETP's who had volunteered to work on the sets of questions. The intent was to develop sets of potential instruments or items that could be used to answer the centralized evaluation questions. Several instruments were created to address these questions. The Core project consolidated all comments and finalized the instruments through careful editing. Ten different sources of data were identified: higher educational institution deans/department chairs, faculty, students, and classrooms; K–12 principals, teachers, students, and classrooms; NSF scholars; and the CETP PIs/evaluators. All of the items on the instruments were matched to the evaluation questions and sources. Instruments included web-based surveys for the higher education deans and faculty members and the PIs; paper and pencil surveys for college students and K–12 principals, teachers, and students; classroom observation protocols with web-based data entry capacity and rubrics for scoring the two classroom artifacts (activity sheets and assessments). All of these instruments were pilot tested and revised based on data from samples of each respondent group. All of the instruments and a crosswalk to the evaluation questions are posted on the website <www.education.umn.edu/CAREI/CETP>.

Figure 1 highlights the essential elements of the Core evaluation. The two triangles illustrate the use of varied data sources at each of the two primary settings. In each setting, data were collected from administrators, teachers, and students as well as external observers. Additionally, artifacts were collected and scored. The frame around the triangles symbolizes that the Core evaluation is only part of the picture. The Core data are “framed,” meaning that they are contextualized, interpreted, and expanded by the unique data from the sites.

Although the Core Evaluation project was required to work with all of the CETP's, individual CETP's were not required to participate in the Core. Therefore incentives were provided to encourage participation. Several face-to-face meetings were held, with the Core or NSF paying consulting time and travel. Individual phone and e-mail contact, as well as an e-mail discussion group, were maintained. In addition, the CETP's were paid for gathering the data needed by the
Core Evaluation project. The staff of the Core entered, cleaned, and specified all data from a CETP and returned the data to it as an SPSS system file so that the CETP could conduct its own analyses with the data. Draft copies of Core Evaluation project reports were provided to the CETPs for comments that were then incorporated into the final report.

The Core Evaluation project used a participatory or collaborative evaluation approach to conduct a “scientific” evaluation. The participatory portion was the involvement of the local evaluators in the specification of the evaluation questions, the evaluation design, the development of instruments, the collection of data, and the development of reports. Burke (1998) suggests that these are the key decision points at which participation is required in order to have a participatory evaluation. The “scientific” portions of the evaluation were the use of standardized instruments and data collection processes, the quasi-experimental, matched pairs design used in the K–12 setting, and the randomization and pre-post assessment of the faculty teaching activities.
CRITIQUE OF THE CORE EVALUATION

The Core Evaluation project highlights many of the complex issues inherent in a process that attempts to bridge the gap between scientific and participatory, collaborative evaluation. This critique of its effectiveness illustrates the strengths and limitations of the attempt. The participation of the CETP evaluators in the definition of evaluation questions resulted in an evaluation that was a mix of “scientific” and other approaches. The CETP evaluators wanted to be able to document what was done as well as the participants’ perceptions of effectiveness. They were interested in questions of process and meaning and wanted the measures directly aligned with what they were doing (Maxwell, 2004). They had goals or standards they wanted the people they worked with to meet. They were much less concerned with finding out if people they didn’t work with obtained these same goals. Along that line, the CETP evaluators were more interested in studying close relationships. The emphasis in the CETPs was on helping faculty and teachers to perform well in their classrooms, and there was consensus as to what constituted good classroom performance. Therefore the CETPs advised the Core Evaluation project members to concentrate on documenting classroom activities rather than student achievement, which was an outcome further removed from CETP programming and one where no consensus existed.

For example, the CETPs were not interested in comparing dean/department chair and faculty perceptions in their institutions to those at other institutions. They wanted the perceptions compared to standards of what might be sufficient perceptual levels. They wanted to determine if a dean’s perception of “substantial” change in teaching efforts in a science department was sufficient to claim program effectiveness. Or at a slightly more “scientific” level, would a statistically significant change in a faculty member’s rating of his/her teaching activities from five years ago to now be sufficient to claim program effectiveness? The CETPs’ interest in documenting effects in people they affected, not in those they hadn’t worked with, was supported by the difficulty in obtaining responses from people who did not work with the CETP. For example, the proposed random selection of faculty members throughout participating institutions was suggested to provide an indication of the effect of the CETP on all of the targeted faculty and the institution as a whole. In reality, the only people from whom the Core project received responses were deans and faculty who were directly involved in the CETP’s programming.
Although the quasi-experimental design in the K–12 setting was “scientific,” it had limitations. One of the flaws was the matching of the CETP and non-CETP teachers instead of random selection. However, because lists of teachers from which to randomly select are not available, this would have been impossible. In practice, although teachers were matched, the matching was not always as comprehensive as it should have been. For example, during the 2003 year of the data collection, surveys were obtained from 287 K–12 teachers, but 49 of them had to be removed because only 238 were teachers with two or three years of experience teaching. Furthermore, there were almost twice as many CETP teachers (165) as non-CETP teachers (73). Another flaw was waiting until the second year of teaching to gather teacher data. This biased the sample in that students who had participated in a CETP program but were not teaching by the second year were not surveyed. This was used, however, because all the sites felt that the first year of teaching would not be representative of how the former student would actually teach. A suggestion for counteracting the flaw would be for the individual CETPs to compare rates over time in terms of graduates obtaining and remaining in teaching jobs.

Triangulation of the different data sources to validate the data obtained was a significant part of the evaluation design. This could be viewed as a strong component, and was philosophically very pluralistic, but it also raises the issue of how data obtained from mixed methods should be combined (Greene & Caracelli, 1997; Lawrenz & Huffman, 2002). It may also cause diffusion of the data collection efforts. Instead of concentrating on obtaining high quality data from one source, high quality data needed to be obtained from several sources. Although theoretically sound, this multiple source approach did not work very well in practice. CETPs were unable to provide complete data from all of the sources. For example, in the 2003 year, one CETP provided K–12 data from 64 teachers and 1076 students but only 11 principals. Additionally, only 17 teachers provided artifacts.

The justifications for using participatory, collaborative evaluation discussed above meant essentially that the data would be more valid because data providers would be more involved. This implies that the local evaluators would be committed to and involved in the collection of Core evaluation data. Unfortunately, this happened only to some degree. The local evaluators were very supportive of the staff of the Core Evaluation project and said they wanted to provide the necessary data. However, as shown in the example above, data
were incomplete. There are several possible reasons for this. As suggested by Straw and Herrell (2002), collaboration must be carefully supported and nurtured. Although the Core project was designed to provide several incentives — interpersonal attachments and collegiality, opportunity for discussion and resolution of problems, efficiency of data collection though the use of pre-prepared instruments, return of collected data in a easily used system file, approval of the funding agency, and money to hire staff to help local sites collect data — the local CETP evaluators may not have seen these as incentives. Not surprisingly, those CETPs that had more positivistic philosophic orientations also had local evaluations more heavily integrated with the Core data and were more likely to provide data than those who had local evaluations that were more interpretive and less dependent on the Core data.

So the combination of “scientific” and participatory evaluation as exemplified by the CETP Core evaluation was only partially successful. The involvement of the CETPs in defining the evaluation questions did make them more supportive of the Core evaluation effort but not supportive enough to provide all of the specified data. This was confounded by the diffusion of the data collection process required by the triangulated design. The data collection at both levels was compromised by the lack of ability of the local evaluators to provide complete and well-matched data. In the K–12 setting the local evaluators were not always able to find second- or third-year teachers directly comparable to the ones who had participated in their CETP. Additionally, even if these comparison-group teachers were located, many were not willing to participate, even with a financial incentive. The lack of matching in the higher education setting made the selection of respondents somewhat easier, but the randomization plan of covering all faculty was unsuccessful.

On the other hand, the data collection did provide a large amount of data over several years from most of the CETPs using consistent instruments. For example, in the 2003 year data were obtained from 16 PIs, 52 deans, 123 faculty, 2,391 college students, 128 principals, 287 teachers, and 4,802 6–12 students. These data could be used within constrained circumstances to provide information about the effectiveness of the program overall and about cross-site patterns and issues. For example, we found that greater percentages of CETP teachers’ classes were viewed as showing accomplished or exemplary instruction than classes of non-CETP teachers, and that students of CETP teachers rated their classes as containing significantly more inquiry-
based activities than students of non-CETP teachers. An additional benefit of the participatory, collaborative approach was that the Core evaluation had standards of behaviour that were agreed upon by the group. One difficulty with experimental designs in general is that they force comparisons. The notion of comparing to a standard is not commonly considered. One approach is seen as better than another, but the absolute value of the approaches on some sort of intrinsic or mastery standard may not be included. For example, CETP teacher classrooms might have more inquiry-oriented science classes than non-CETP teacher classrooms, but they both might be well below the desired level of inquiry. The design in the Core used standards of behaviour as well as comparisons to assess program effectiveness.

The Core evaluation process helped to define the essence of the program from the sites’ perspectives. The limitation of this was that the CETPs had ideas of the essence and valuable outcomes of the program different from those that could be ascertained using the “scientific” evaluation favored by NSF. On the other hand, by agreeing on central questions and sharing the work involved in developing instruments and processes, individual sites were freer to concentrate on the unique aspects of their approaches. The Core evaluation also helped to provide the objectivity hoped for in a “scientific” evaluation. The Core team independently rated the lesson plans and assessments and provided quality control through the standardization of data collection processes and instruments. For example, the Core team, with the help of the CETPs, developed a video training program to help ensure reliability of classroom observations. Furthermore, the Core team provided data collection, input, and analysis services for all, thereby providing an economy of scale. Additionally, the existence of the consensus represented by the Core allowed the CETPs to present a more consolidated front in negotiations with the funder. The Core team also facilitated the sharing of expertise across the sites (Leff & Mulkern, 2002).

Other types of experimental designs could be implemented using the core concept that might eliminate some of the problems experienced by this Core Evaluation project. The heavy use of triangulated data could be replaced by the judicious selection of one source for each type of data. To combat the difficulty of obtaining data from comparison groups, some sites could implement first while the others served as comparison sites. As the different sites entered the study, time effects could be estimated. In other words, exposing subjects to the treatment for different lengths of time should produce
more significant effects. Alternatively, different aspects of the program could be tried at some sites and not at others and the effects compared. Results from testing one approach at some of the sites could be used to improve it, and then the approach could be tried out again in different sites. The use of participating sites as comparison groups helps to ameliorate the difficulties involved in finding comparison sites and also helps to control for selection bias because all of the sites were interested in the program and all were “good” enough to get funding. Additionally, involving all of the sites in the experiments would help to sensitize them to the procedures and approaches being used and how to improve them.

SUMMARY

The view of “scientific” research and evaluation put forth by the U.S. Department of Education asserts that experiments provide the best data regarding effectiveness of programs. The CETP Core Evaluation project was an attempt to provide a “scientific” evaluation but through a participatory approach. A collaborative, consensus-building approach was required because the CETP projects had already been funded and could not be required to participate in an evaluation. Using a participatory, collaborative evaluation approach provides many advantages, but the CETP Core evaluation demonstrates that it is not necessarily the best way to provide “scientific” evaluation. The inclusion of all of the local evaluators and their philosophic perspectives resulted in an evaluation that combined many different evaluation models and had a design incorporating “scientific” and descriptive approaches. The local CETP evaluators generally had more subjectivist-pluralist philosophical stances, which moved the design into a more descriptive realm where the perspectives of different subgroups and contexts were important. This stance was consistent with the philosophies of the CETPs, and perhaps social service programs in general, which were designed to provide excellent preservice teacher education to diverse students and to produce teachers committed to providing classroom instruction that would allow all children to succeed in science and mathematics.

The objectivist-utilitarian philosophy base of the “scientific” approach to evaluation was less appealing. The strong relationships built through personal contact and professional respect among the CETPs and the Core Evaluation staff resulted in the development of instruments and the collection of data that would not have been possible without the participatory approach. Despite these relationships,
the data collection was incomplete. More descriptive data directly related to the CETPs was provided to the Core perhaps because these data were seen as more valuable. The “scientific” portions of the evaluation design were apparently less valued, as less data were provided for those aspects of the evaluation.

The lessons learned by the Core Evaluation project can help to inform other groups planning “scientific” evaluations. It appears that many evaluators may not be supportive of a completely “scientific” approach. Therefore it seems prudent to design “scientific” evaluations that also include other evaluation approaches. In particular, methods that gather information on local contextual variations and differential effects on various subgroups should be included. This speaks for mixed methods approaches as discussed by several evaluation researchers (Caracelli & Greene, 1997; Greene & Caracelli, 1997; Lawrenz & Huffman, 2002). Special attention to the philosophical underpinnings of the approaches, however, appears to be called for. Even though the local evaluators agreed in principle with the “scientific” parts of the evaluation design, these data were less complete. This result argues that significant amounts of time need to be spent constructing understanding of and commitment to a “scientific” design. It is also apparent that for data collection to be complete, whatever data the core evaluation needs must be completely integrated into the local evaluation plan.

It is also critical that if a participatory, collaborative approach is used, the evaluation design cannot be specified before the participation. To have true participation and commitment, the evaluation must be subject to consensus and revision by the group at several decision points (Burke, 1998). This means that there cannot be a predetermined and unchangeable “scientific” design. The blend of these two, which was exemplified by the Core evaluation, produced an evaluation that was only somewhat “scientific.” If a completely “scientific” design is critical, it is probably necessary that the participatory element be de-emphasized to perhaps providing advice and suggestions, and that the provision of data to the scientific study be required by the funder. This type of external, top-down data collection, however, also raises problems. How to combine participatory and multi-site evaluation approaches is an important topic for further discussion (Lawrenz & Huffman, 2003).

The local CETPs clearly felt that relying upon only quasi-experiments to evaluate their programs would provide incomplete under-
standing. Certainly their view was that to fully understand these complex social and educational programs, it was essential to employ a variety of methods. Approaches that merge methods, such as the Core evaluation, are valuable. The core evaluation concept provides a way to assist local sites in their own evaluation efforts and to provide aggregateable data that may be used in experimental or quasi-experimental designs. The Core approach provided data collection procedures and instruments grounded in the reality of the program as it was implemented, while at the same time producing an external, objective examination of the program. This approach to evaluation helps to meet the requirements of objective, scientific evaluation while, at the same time, valuing and incorporating local evaluation efforts.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundations grant DUE 9908902.

REFERENCES


Frances Lawrenz is a professor in the Department of Educational Psychology at the University of Minnesota, where she teaches classes on research design, methodology, and evaluation. She has a background in chemistry and mathematics and has worked in science program evaluation for over 20 years, including two rotations at NSF as an evaluation specialist. She has been principal investigator on numerous science and mathematics evaluation projects. She has served as the assistant vice president for research for the university, and as the chair of the Department of Educational Psychology in the College of Education and Human Development. She is currently the Wallace Professor of Teaching and Learning and recently received the All-University award for outstanding contributions to graduate education.

Douglas Huffman is an associate professor of science education at the University of Kansas. He has a background in civil engineering, science education, and program evaluation. He has served as an external evaluator on numerous NSF-funded projects including the Collaborative for Excellence in Science and Mathematics Teacher Preparation (CETP), an evaluation of State Systemic Initiatives (SSI), and several curriculum development and teacher enhancement projects. He currently serves on the editorial board of the Journal of Research in Science Teaching and has published more than 20 peer-reviewed articles as well as numerous evaluation reports.