WHEN (NOT IF)
EVALUATION
FLEXIBILITY IS
DESIRABLE
Examples From the
CPHPE Initiative

LARRY HEMBROFF
HARRY PERLSTADT
REBECCA C. HENRY
ANDREW J. HOGAN
CAROL S. WEISSELT
Michigan State University
CAROLE J. BLAND
University of Minnesota
DONA L. HARRIS
East Carolina University
School of Medicine
JACK H. KNOTT
University of Illinois–Urbana
SANDRA M. STARNAMAN
Colorado Mountain College

The evaluation literature often debates whether evaluators should be flexible in evaluation design and activities in order to collaborate with program directors and be responsive to programming needs. Two conditions are specified under which evaluation flexibility is not only desirable but essential. Two examples from the cluster evaluation of the W. K. Kellogg Foundation’s Community Partnerships for Health Professions Education initiative are provided to illustrate why flexibility under these conditions proved to be essential. One of the examples, related to the “community” involvement in the initiative, illustrates the need for flexibility as programs experience goals clarification. The other example, related to the coincidental national health care reform efforts, illustrates the need for flexibility both to capture programs’ efforts to protect their integrity and to ensure against spurious conclusions as a result of external turbulence in policy environments. How the cluster evaluation team addressed these issues is also described.

AUTHORS’ NOTE: Address correspondence to Larry A. Hembroff, Ph.D., Institute for Public Policy and Social Research, Michigan State University, 321 Berkey Hall, East Lansing, MI 48824; phone: (517) 355-6672 ext. 122; fax: (517) 432-1533; e-mail: Larry.Hembroff@ssc.msu.edu.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 22 No. 3, September 1999 325-341
© 1999 Sage Publications, Inc.
Scriven and Patton have staked out positions on largely opposite sides of the debate regarding the proper role of evaluation and evaluators in relation to the programs they evaluate—"distance vs. transactional evaluation" (Scriven, 1977) and "goal-based vs. goal-free evaluations" (Patton, 1997). Patton supports a more flexible, intervention-oriented evaluation to the extent that it supports and reinforces accomplishing desired program goals. He recognizes, nonetheless, that there are limits to flexibility and intervention beyond which an evaluation may lose its credibility and utility. Scriven sees such flexibility and intervention as potentially serious breaches from evaluation objectivity and threatening the likely validity of analyses regarding program outcomes.

In its present form, we believe the Scriven-Patton debate is mispecified. The issue is not whether evaluation flexibility or intervention is or is not acceptable/desirable. The correctly specified question is, Under what conditions is flexibility or intervention acceptable/desirable and under what conditions is it not?

In fact, many others have also rejected the absolutism represented by these two extreme positions. The W. K. Kellogg Foundation, for example, has attempted to develop a more pragmatic approach (Millet, 1996). The foundation expects evaluations to help improve projects, to strengthen foundation programming, and to inform policy making. Similarly, evaluators involved in the Center for Substance Abuse Prevention Community Partnerships have moved from an internal research evaluation toward "empowerment evaluation," which includes collaboration among coalition members, staff and evaluators, improving the coalition and its plan of action, and flexibility in solving problems and trying new approaches (Butterfoss, Goodman, Wanderman, Valois, & Chaisson, 1996).

Most evaluators recognize that evaluations tend to change their focus over time and the literature contains some examples (Dugan, 1996; Jacobs, 1988; Trippi, Felin, & Epstein, 1971). These examples, however, link the changes in evaluation to stages of program development or difficulties in program implementation. These are two conditions under which evaluation flexibility is actually desirable.

In this article, we try to identify and illustrate at least two more conditions under which flexibility is desirable and may lead to stronger evaluations than would otherwise occur. These other conditions have to do with program goals. One involves the related clarification of program goals and the second involves external social changes that also affect program goals. Our illustrative examples are drawn from our work on the multisite evaluation of the Community Partnerships for Health Professions Education (CPHPE) initiative funded by the W. K. Kellogg Foundation (WKFP). We will describe the initiative, the evaluation approach, and two detailed examples that illustrate why evaluation flexibility was essential for a meaningful evaluation of the program.

THE COMMUNITY PARTNERSHIPS FOR HEALTH PROFessions EDUCATION INITIATIVE

In the late 1980s, given their interest in systems change and concerns about national health care, program leaders at WKFP recognized the pivotal role health professions education plays in the development and distribution of health care practitioners. The WKFP concluded that there were too few suitably trained primary care health professionals and that those then in practice were misdistributed, resulting in significant populations of underserved Americans around the country. The foundation attributed much of this to the increasing frequency of health professions students choosing specialty careers over primary care rather than to an absolute shortage of health professionals.

In WKFP's view, health professions education had become overly dependent on hospital-based venues as the clinical settings in which students are trained and professional values slurred. They believed that the experiences in teaching hospitals promote specialization, rather than primary care or multiprofessional team approaches (Kendall, 1961, 1971; Perlstein, 1972a, 1972b).

To change the outcomes—that is, to increase the numbers of health professions students choosing primary care and choosing to practice in underserved areas—the foundation required demonstration project sites to change in three specific ways. They had to change their health professions curriculum, to move much of the clinical training out of
hospitals and into academic community health centers, and to create new community-university partnerships to oversee the new curriculum and clinical sites.

Weiss (1995) contends that evaluations work best with programs that are based on explicit or implicit theories about how and why the program will work. The articulation of the program's underlying theory of action and change enables an evaluation team to adopt a theory-driven approach (Chen & Rotai, 1983, 1989). The fact of the matter is, however, that much of the "explicitness" of the model underlying the CPHPE's theory of change and theory of action is the result of the evaluation team's efforts to explicate that model so it could construct a meaningful evaluation rather than what was explicitly articulated in the program's planning documents. This is actually all too common.

The foundation's theory of action for CPHPE involved an outside-in strategy for systems change. This change strategy was based on four assumptions. First, there had to be pressure from the outside to motivate academic institutions to change the way they do their business. The size of the award, 6 million dollars, captured the attention of the institutions. Second, to sustain desired change, the system supporting the status quo had to be altered. Policies at the institution, state, and national level had to be examined for their potential in sustaining change. Third, a new organizational structure had to be established at each site that would give the community a stronger voice in establishing new goals for health professions education. Finally, CPHPE sites had to move much of the clinical components of their curriculum outside of hospitals. These four assumptions about change gave direction to the evaluation team as it developed a model for program evaluation.

In 1991, WKRFP awarded approximately 6 million dollars during a 5-year period to each of seven community partnerships around the country. Each partnership joined medical, nursing, and other health care professional schools together with lay community representatives. The seven new partnerships were located in Georgia, Hawaii, Massachusetts, Michigan, Tennessee, Texas, and West Virginia. The settings were varied: large metropolitan areas, smaller cities, and rural regions.

In addition, the foundation funded a multisite or cluster evaluation of the program, an approach to evaluations across multiple similar projects developed by the evaluation unit of the foundation (Kellogg Foundation, 1993, 1996; Sanders, 1997). A cluster evaluation is intended to assess the progress, outcomes, and impact of a programming initiative by looking across a group of similar projects and their varied contexts and approaches to identify common patterns and themes to learn what happened and why. Compared with other types of evaluations, it is a somewhat more collaborative effort between the funding agency and the evaluators. An important reason for this is to provide information for and about programming directives so as to minimize the opportunity for program failure. Therefore, the evaluation must meet the criteria for a sound evaluation while being flexibly responsive to the needs of the funding agency, the projects, and the target populations.

PROGRAM GOALS AND EVALUATION CONSEQUENCES

The CPHPE had multiple goals. Ultimately, the initiative hoped to increase the numbers of suitably prepared primary care practitioners, which would improve health care and health in formerly underserved areas. It also had multiple intermediate outcomes, designed to facilitate the long-term impact. Creating a partnership organizational structure, moving education to community settings, changing the curriculum, and implementing a public policy plan for project sustainability were all important intermediate goals targeted by the evaluation. Figure 1 illustrates some of the major components of the program and factors of concern in the evaluation of it.

Because the program was intended to create change in students via change in academic institutions (i.e., curriculum, faculty, and university policies), communities (e.g., partnership organizations), and public policies, an evaluation design was developed to look at these areas of change (institutions, community, and public policy) in terms of the projects' outcomes, contexts, and implementations (Kellogg Foundation, 1993, 1996). The evaluation called for a longitudinal design that included multiple methods and instruments of data collection, in
many cases the evaluation plan permitted at least two observations in each of the three arenas of change.

Table 1 lists the major evaluation activities that were planned originally and that were actually implemented. The table indicates that a number of activities that were included in the original evaluation plan were eventually dropped, whereas a number of other activities were added that had not been planned originally. The final cluster evaluation data collection included site visits, annual meetings with project evaluators, annual reports from each site, personal interviews with key faculty, staff, and community representatives at each of the projects, surveys of the faculty and students at the participating institutions, and reviews of relevant project documents that were all a part of the original plan.

The faculty and student surveys were designed to make observations related to the institution/curriculum areas of change. The document reviews were primarily designed to focus on policy issues within the institutions, the professions, and the state and national legal and budget environments. The annual reports and site visits addressed questions related to all three arenas. The policy makers survey, the partnership participants survey, and the leadership study were added or significantly modified with respect to timing as a result of program adjustments on the part of the foundation. Two examples of these will be discussed at length below. As a result of these program adjustments, several of the other evaluation data collection efforts were dropped from what had been included in the original plan. These will also be indicated in the discussion below.

## Change and Flexibility

The original evaluation plans were developed prior to the formal beginning of any of the projects. Consequently, the evaluation plans were based on the proposals submitted by the individual projects and the goals and design of the CPHE initiative itself as described by WKKF. As is often the case, "mission creep" during program planning stages had a tendency to make grand, overly ambitious goals.

Such grandiose program visions are often moderated significantly...
when confronted with actual implementation and context. When that happens, evaluations must be flexible enough to be scaled back as well, or they will, in fact, evaluate programs that were never actually implemented.

We have indicated that there are conditions under which evaluations need to be flexible as well as theory driven or goal focused. We will now present two examples from this project to illustrate two conditions that require flexibility to make the final evaluation meaningful. We will discuss how the cluster evaluation team modified its strategy within the general evaluation model to assess the issues relevant to the program as it was actually implemented. One of the examples comes from the community arena and the other comes from the policy arena.

Example 1: Community. The CPHPE program’s theory of action held that health professions students would find primary care in community-based settings attractive and would then choose community-based primary care as a career, ultimately contributing to the goal of increasing the supply of primary care providers in medically underserved areas and improving the communities’ health status. The program’s theory of change called for community members to participate in curriculum change and to enrich the community-based experience. They would inform the curriculum developers about the specific needs and concerns of their community and offer to provide opportunities for students to become immersed in local health activities and community events. The community-based health centers, then, would provide a credible alternative to hospital-based training by bringing university-based teaching and scholarship out to the community. These centers would also expand professional socialization experiences outside of the university hospital by involving local providers and community leaders in students’ learning experiences. Accordingly, the original evaluation called for developing community health profiles and community health service and resource inventories and for conducting community health status and needs assessment surveys. These would serve as baselines against which to compare postprogram assessment at the community level. The size of the discrepancy between needs and community resources would provide the measure and shape of the communities’ unmet needs at start-up. The comparison of the postprogram with baseline surveys and profiles would indicate the relative success of the program at achieving its goal.

However, the initiative had barely begun when it became obvious to both programmers at the foundation and the cluster evaluators how unreasonable it would be to rely on the measurement of these aspects of the initiative’s goals as a basis for judging the program’s overall outcomes. Because of the lengthy training phase for health professionals (especially medical students who would then go through residency programs before entering practice), it was beyond the scope of the evaluation to monitor this long-term outcome. Even if all the health professions students ultimately chose primary care and chose to practice in underserved communities, it was unreasonable to expect that they would settle in the same communities in which they trained.

In addition, the costs and disadvantages of this approach were quite obvious. Doing needs assessment and health status/utilization surveys of residents of the nearly 50 communities involved in CPHPE would be enormously expensive and would undoubtedly drain resources away from other evaluation activities or would require supplemental funding from the foundation. In addition, because medical students would not be choosing their practice locations until the end of their residencies, the posttest measurement of impact on the communities’ health would have to wait until these students entered practice some 10 years after the 5-year project began.

The evaluation team met with the foundation to present five alternative strategies for assessing more realistically measurable program outcomes in the community arena. The result was a substantial clarification of goals and outcomes. The discussion was especially instrumental for the evaluators and for the foundation in identifying what outcomes in the community arenas would be key indicators of success. The foundation favored narrowing the community evaluation to focus on the incipient partnership boards and project committees whose university and community people were supposed to share program decision making regarding curriculum, clinical experiences, research, and outreach.

The goals horizon for the community arena was shortened from the long-term outcomes of the unmet needs of an underserved popu-
lution to the more intermediate term development of community-empowered partnership structures. Consequently, the cluster team set aside the planned health status and utilization surveys and developed methods to examine community empowerment. Site visits, interviews, and focus groups were added to examine their commitment to, and participation in, partnership activities and decision-making; which were represented, and how well the partnership's goals met community needs. Previous research (Bliss, Zonta, Sturman, & Rosenberg, 1992) indicated that particular leadership structures and approaches were more likely to result in successful organizations in groups such as these; therefore, respondents were asked to assess the behaviors and approaches of partnership leaders. In turn, leaders were asked to assess their cognitive frames as they perceived the partnership organizations (Bolman & Deal, 1994).

The impact of these evaluation activities on decision making, however, was significant. In all cases, the initial partnership boards were dominated by the health professions schools in numbers and leadership, so that the way of doing business took on a decidedly academic flavor. The early site visits revealed that almost all the community representatives were unprepared to participate effectively in curriculum development or even to present their community's unmet needs. The presence of lay community members on the curriculum committees, however, did improve faculty interest and on their best behavior as they hampered out the program's curriculum changes. The community members did mobilize community support for the program and organized orientation and social activities to better acquire the students with the communities.

The evaluation team included specific questions on the participants' survey to examine community participation and empowerment. The evaluation team presented the survey findings at a CPHEP network meeting attended by representatives from all projects. By that time, community representatives were a majority on three of the seven partnership boards. The results of the survey showed that members of boards with community majorities had a more open decision-making process, shared information, and were more satisfied with their participation and the projects' accomplishments.

As a result of this monitoring and reporting on the effects of board composition on communities' voice in decision making on the partnership board, other sites felt compelled, with foundation encouragement, to establish community majorities on their boards as well. By the end of the 3rd year, all but one of the seven partnerships had a community majority. This shift in program implementation strategy led the cluster evaluation to question the impact of this or least technical composition change on actual board function. As a result, the cluster team requested and received approval to convene one additional (and last) meeting of project leaders representing academe and community from each of the seven projects. In that meeting, members of the evaluation team conducted focus groups for the project leaders to summarize their experience for WKKF. However, judging from the subsequent site visits and the leadership focus group held at the end of the project, those sites that initially, hastily added community representatives to achieve a numerical community majority still had not really empowered their community members.

As a result of the evaluation effort's shifting to capture the flow in programming directions, the team was able, by the end of the initiative, to offer recommendations on effective configurations for university-community partnership organizations for future foundation programming. Developing these was, in fact, the central outcome goal with respect to communities of the CPHEP initiative from the foundation's point of view. But it was only through the discussion with the program director regarding the community portion of the evaluation that the evaluators ultimately understood the priority of the organizational structures to program goals. This was not previously clear or fully appreciated in the original evaluation plan.

Thus, to evaluate the program based on its underlying model of changes and its vague conceptions of community, the evaluation approach had to be flexible enough to provide feedback that led to readjustments in both program emphasis and evaluation efforts. In particular, the outcome evaluation was modified from health status to community empowerment within the partnership organization, which
was a much closer fit to the program's theory of action. It is interesting to note, however, that although the evaluation findings had a direct impact on both foundation programming and project structure, changes that were made in response to the findings only modified form and not substance or functioning.

Example 2: Public policy. The second example comes from the evaluation in the public policy arena. In 1991, at the beginning of CPHPE, a major focus of the policy arena was health professions education funding policies, practice regulations, and accreditation standards affecting the sustainability of the CPHPE approach. The evaluators conducted an inventory of relevant state policies that would be either favorable or unfavorable to the multidisciplinary community health centers and other out-of-hospital settings in which the project would be implemented (Weissert, Knott, & Strecher, 1994). Such policies were expected to affect student career choices as well as affect the long-term sustainability of the partnerships.

The foundation encouraged projects to inform state officials and some representatives to Congress so they understood the importance of community health centers in training students for primary care careers. The goal was for each project to identify policies that could provide financial sustainability to community-based health professions education and, through consultation, testimony, and lobbying to stimulate or support policy changes that would be supportive of these innovative educational partnership structures. Accordingly, the original evaluation plan called for what was essentially to be a pretest-posttest measurement of the relevant policies, regulations, and standards to see what had changed; to examine legislation professed; and to track contacts, testimony, and news reports of the projects related to these relevant policies, regulations, and standards.

When Clinton was elected president, there was a profound shift in attention on the part of the federal government and the foundation to national public policies concerning health care organization and funding. Although policy had always been central to the evaluation, the importance of riding the crest of the political wave significantly accelerated program activities in the public policy arena and shifted the focus from state to national health policies. This seemed especially crucial given the prominence primary care and managed care were receiving in the health care reform discussions that surfaced publicly.

The foundation, convinced of the wisdom of its program conceptually, was also fearful that unprepared policy makers might drown the fledgling partnership model with reforms that would create health care and health care financing countercurrents. The foundation shifted its focus and planned schedule. It chose to educate policy makers and to direct the projects to participate in the efforts with it and in addition to it. Consequently, the evaluation team added a survey of key Washington health policy makers to the evaluation plan.

To assess the impact of the foundation's efforts at informing the national health care reform debate, the cluster evaluation team conducted a series of interviews with more than 100 key health care policy makers from Congress, interest groups, and the executive branch. The purpose was to find out if they were familiar with the CPHPE projects and whether the information about the projects was shaping any specific elements of legislation under consideration. Because other foundations were also active with related initiatives, the interview also asked similar questions of those foundations' efforts so that the relative effectiveness of the foundation's information could be evaluated (see Weissert & Knott, 1995). The cluster evaluators also interviewed a number of Washington observers and CPHPE contractors to understand the process the foundation used to inform policy makers.

Midway through the CPHPE, national health care reform died. As a result, the attention shifted once again to the state-level concerns. The evaluation strategy regarding the policy arena had to adapt in order to assess the outcomes of these shifting political tides. Developing an inventory of the relevant policies favorable or unfavorable to sustaining community-based academic health centers as teaching venues remained an appropriate evaluation strategy for assessing the policy context. A comparison of state policy contexts at the beginning with those at the end, combined with a review of documents regarding project policy-related efforts could still be used to understand change. However, this could not capture the effects of the foundation's or the projects' policy-informing efforts. Consequently, the evaluation approach had to be modified again. As a result, a similar survey of state-level policy makers was also added to the evaluation effort.
Thus, to understand the changes in the contexts in which the projects were being implemented and to evaluate the efforts on behalf of the program and the projects to effect favorable changes in those policy contexts, the evaluation had to change. And, because the times turned several times, the evaluation strategy had to change several times or risk missing most of the real action. Flexibility in the evaluation design was absolutely essential.

**DISCUSSION AND IMPLICATIONS**

We began by suggesting that the argument about whether evaluation should be flexible or not, whether evaluators should be distant from the programming or not, or whether evaluations should strictly focus on program goals or not is misspecified. Rather, we suggested, whether an evaluation should be flexible or not is conditional.

We have identified two sets of conditions under which, we believe, flexibility was not only acceptable but was necessary for the evaluation to be meaningful. The first is from our community arena example. When evaluators are brought in (as most of us think they should be) prior to actual program implementation, and where the goals of the program are many and are as yet untempered by the realities of actual implementation, evaluation flexibility is necessary to that the final evaluation strategy takes into account the inevitable goals clarification of the program as carried out. In this case, it is also highly desirable and illuminating to document the goals clarification so that the evaluation does not come to declare the program a success simply as an artifact of having clarified away all goals.

The second is from our policy arena example. This is a set of conditions that occurs when events exogenous to the program’s model but intimately related to the program’s goals create powerful perturbations in the context in which the program is being implemented. Evaluation flexibility so that efforts to protect the program from these turbulent forces can be assessed are essential to sorting out whether a program fails because of a flaw in theory or because it was overwhelmed by more powerful forces. Or, evaluation flexibility in this case is necessary to be able to disentangle the impact of these forces from those of the program itself when trying to discern the impact of the program. In the example we gave, responding to the external forces necessarily redirected, albeit temporarily, resources of the programs and program directors from the implementation activities they had planned. Being flexible enough to capture that redirection of resources is essential so that it can be properly taken into account when assessing the program’s effectiveness.

In the examples we have offered, evaluation flexibility resulted in redefining the focus of the evaluation in two arenas, reconfiguring data collection strategies, methods and timing, and becoming more directly involved in program interventions. Evaluation flexibility had its disadvantages as well—financial and otherwise. The shifting focus of the evaluation microscope meant that the opportunity to collect some time-series data was lost. It also meant that new instruments had to be developed to examine concepts relevant in the new/shifted focus of activity rather than being able to reuse, at some postintervention points, instruments developed to measure concepts at baseline.

In our case, the shift to partnership organizational structures and away from communities of people was disappointing, but it did reveal yet another high-priority goal/assumption of the initiative. Examining the affective, social, as well as health consequences of establishing such partnerships and community-based health centers would have been fascinating and a potentially rich source of data on many counts, but it would no longer have been as relevant to evaluating the program as it evolved. Intensifying the focus on policy change because the potential opportunity to influence change increased resulted in allocating far more site visit resources to policy issues than team members concerned with the community and institutional curricula would have preferred. Some data regarding these other arenas undoubtedly were lost as a result.

Nevertheless, under the conditions we have identified, evaluation flexibility has enormous strength to the extent it enables the evaluators to examine the program as it evolves. It enables the evaluators to take advantage of opportunities that arise that they could not have anticipated initially and might have missed if their design and strategy had been dipted in bronze at the beginning of the program. A static
evaluation design may be methodologically or intellectually pure, but risks being practically irrelevant.

NOTES

1. The Community Partnership for Health Education (CPHE) was one of the largest (in terms of dollars) W. K. Kellogg Foundation initiatives to use a chart evaluation. Chart evaluation, like this program and its evaluations, have been described in length by Robinson (1996).


3. One of the authors was evaluated a personal care intervention program for which the project designers had obtained the impact of the project would be for a model to provide care and deliver better for more than 500 pregnant women in a 2-year period. In actual implementation, these goals were confirmed with the reality that standards of care limit the number to only 8 to 10 evaluations per month, or a maximum of 100 for the 2 years.

4. In this case, the criteria path was condition is not sufficiently met to allow the evaluators to determine if program success or failure was the result of inadequate intake.

REFERENCES


