Pergamon

Children and Youth Services Review, Vol. 19, No. 7, pp. 607–614, 1997

0190-7409/97 $17.00 + .00

PII S0190-7409(97)00048-0

# Epilogue: Four Lessons from Evaluating Controversial Programs

**James Knickman**
**Paul Jellinek**
The Robert Wood Johnson Foundation

The papers in the final section of this volume describe in great detail the major methods of program evaluation. In this epilogue, we discuss how we grappled with various evaluation design issues within the context of one demonstration program. We conclude by sharing four lessons we learned along the way, lessons that we believe will be useful to those who evaluate programs as well as to consumers of such evaluations.

Over the past 25 years, the Robert Wood Johnson Foundation, a national philanthropy dedicated to improving the health and health care of Americans, has supported a wide range of demonstrations directed at changing the health-related behaviors of individuals (Robert Wood Johnson Foundation, 1997; Isaacs & Knickman, 1997). Thus, foundation staff are constantly faced with the challenge of deciding when an evaluation is feasible and worthwhile and how the real-world constraints related to project implementation interact with the research design requirements of legitimate evaluation.

In principle and by tradition, the foundation is dedicated to objective, sophisticated evaluations of its demonstration programs. A principal aim of the foundation's grant making is to try new strategies for addressing a range of health-related problems and to study whether these new strategies achieve the foundation's stated goals. In many cases, a formal evaluation, grounded in methods of social science research, is seen as an important element of our attempts to foster innovation. A formal evaluation is often the best means of developing "proof" that a new idea is worth replicat-

Reprints may be obtained from James Knickman, Ph.D., Vice President, Robert Wood Johnson Foundation, P.O. Box 2316, Princeton, New Jersey 08543 [jrk@rwjf.org]

ing—often with the support of the government or other sponsors—on a national basis.

At times the requirements of evaluation have gotten in the way of trying out new ideas, and at times it has not been the purpose of a demonstration to "test a hypothesis" about a well-defined "stimulus" expected to lead to a response. Hence, the evaluation imperative is balanced by broader goals of the foundation's activities: to make sure that we learn from what we do and, most important, to help foster social change that improves the health and health care of citizens.

Too often, however, researchers and consumers of evaluation data have overly high expectations for evaluations. They need to appreciate and respect the constraints facing the other side of the demonstration process.

## The School-Based Health Centers Demonstration

While no single case can raise all the issues that need to be considered when implementing field demonstrations, our experience in implementing a 19-site demonstration of high school and middle school health clinics that the foundation supported in the late 1980s offers a useful context (School-Based Adolescent Health Care Program, 1994; Lear et al., 1991). At the time, school-based clinics were seen as a new approach to improving access to basic health care services for low-income youth, and the clinics were considered a potential locus for efforts to affect high-risk behavior of youth, such as sexual activity and drug use. From the start, the foundation was interested in both potential goals of school-based health clinics.

When the initiative was designed, however, the issue of counseling about the use of condoms or distributing condoms or prescribing birth control devices was highly controversial among the public—as it remains today. *Nightline* aired a widely discussed episode on the issue in the fall of 1984, and newspapers printed detailed accounts of local and national debates (often heated) on the issue. One side argued that the rights of parents to control the discussion of norms for sexual behavior with their children were paramount. The other side countered that concerns about HIV (the virus that causes AIDS) infection and other sexually transmitted diseases, combined with the growing prevalence of teen pregnancies, were grounds

for a school-based program (Council on Scientific Affairs, 1993; Miller, Card, Paikoff, & Peterson, 1992).

The foundation approved the clinics before an evaluation was discussed. While the foundation had planned a formal evaluation, other considerations were more important. Because the clinics had the potential to spark controversy, the program staff focused most attention on designing an initiative that could be implemented in local communities. The primary question for program planners was whether school-based clinics were viable on a broad scale: would diverse school districts take the risks necessary to get clinics up and running?

Once the project did get underway and the 19 communities made the attempt to start clinics, the foundation staff became more interested in a formal evaluation to test whether use of health care services increased for students with access to a clinic and whether clinic use reduced risky behaviors, such as early initiation of sexual activity and substance abuse.

As Kisker (1997) notes in this volume, a random-assignment design was not possible since schools had already been selected (nonrandomly) without agreeing to the rigors of randomization. Moreover, the foundation did not want to make the clinics available only to randomly selected students and to withhold the services from others.

The original evaluation design that was approved and funded included a matched comparison sample of schools not operating health clinics in the same school districts as the treatment sites. It was considered crucial to select comparison schools in the same school district to hold constant a range of environmental factors that could influence outcomes and to obtain better cooperation from schools in a district with a direct, active stake in the demonstration.

At this point, the constraints of program implementation and evaluation requirements collided. Senior foundation leaders became concerned that the adverse reactions to students being surveyed about sexual behavior in comparison schools could lead to parental backlash and erode the much-needed support of the participating school districts to implement the clinics. The negative community response to a youth survey in Baltimore (not funded by the foundation) sensitized the leadership to the potential risk that such surveys could pose to the program. Since the primary evaluation question had always been whether this initiative could get up and running, it was decided to exclude comparison schools from the evaluation's surveys of student behavior.

We considered abandoning a formal evaluation completely but instead decided on a strategy that would compare changes in high-risk behaviors among students in the schools with clinics to a national sample of urban youths. The urban youths, contacted in a random-digit telephone survey in a sample of cities, comprised a "reference sample." We compared surveys of both groups to determine how behaviors changed over time.

In the case of high-risk behaviors, the findings were quite difficult to interpret. Such behaviors became more frequent over time, as is always the case as high school students get older, and the prevalence of risky behaviors was significantly higher in treatment schools at baseline than for any reference samples. Since rates of high-risk behaviors tend to plateau in most student populations as youths get to be 18 years old, and since baseline rates were not comparable, it was difficult to learn much from the observed changes in rates over time.

Later, however, the evaluation's expert advisory committee voiced significant concerns about the reference sample as a legitimate comparison point for two key reasons: the treatment schools were not at all representative of the typical urban school nationally, and there was no information about any health education programs to which the students in the reference sample had already been exposed. Thus, it was impossible to know what "treatments" the reference sample had received.

Nevertheless, the findings from the process evaluation were intriguing. There were clear signs that access to the clinics increased the use of basic health care: rates for a range of preventive, diagnostic, and primary care increased over time between baseline and later years, and data from a range of other youth samples showed no comparable utilization increases (Kisker & Brown, 1996).

These utilization rates led us to attempt another strategy for assessing the impact of the program: "dose response." Treatment schools with very active health clinics—as judged from qualitative site visit information—were compared with treatment schools with less active health clinics. According to this analysis, the more active clinics had student bodies that were more sexually active and that had larger increases in sexual activity over time. The validity of these findings, however, was also questionable, as the evaluation team emphasized. The background characteristics of students in the low-dose and high-dose schools differed significantly, and baseline prevalence of high-risk behavior differed significantly at baseline across the two sets of schools. Also, most of the differences in high-

activity and low-activity schools centered around differences in how much basic health care was available and not in how much counseling about high-risk behavior was available (Kisker & Brown, 1996; Marks & Marzke, 1993).

Carefully selected comparison schools would have helped tremendously in understanding the patterns of findings. As it turned out, neither the surveys nor the program engendered much controversy during the course of the initiative. As one program participant noted, "We were fighting the ghost of the past in designing the evaluation."

## Four Lessons on Evaluating Controversial Programs

The foundation's experience in designing, implementing, and evaluating programs such as school-based clinics has led us to conclude that there are four key lessons for evaluating controversial programs.

*Lesson 1: Avoid experimental evaluations if the initiative's main goal is program viability.*

As Devaney (1997) notes elsewhere in this volume, there are process evaluations and outcome evaluations. In the example of school-based clinics, the foundation's main interest was getting clinics into schools and making them operational. The requirements of a rigorous evaluation—finding control or comparison schools and asking students detailed questions about high-risk behaviors—caused concern that participants would abandon the initiative. The imperative to get a program up and going will always rule the day and overshadow evaluation considerations when viability is in question.

Process evaluation also enables individual sites to vary the treatment. This is important if the planners believe that by allowing local sites some leeway on the elements or characteristics of the initiative, the chances for local success at implementation will increase. A process evaluation, of course, will not indicate whether the intervention affected behaviors, which is often a central question.

*Lesson 2: Not every initiative has a doable-outcome evaluation.*

This lesson should be self-evident, but at times the momentum to do an outcome evaluation is so strong that simple lessons are forgotten. Often the desire to learn what "works" moves an evaluation forward. A rigorous evaluation is crucial if the goal is to test hypotheses about the relationship between a program and an outcome (Campbell and Stanley, 1966). If legitimate comparison information is not available, however, an outcome evaluation is not worth the financial investment. Sometimes an outcome evaluation is impossible because the scale of an intervention is too small. For other programs, insufficient time for their effects to set in, and the lack of other "necessary" conditions for positive outcomes, makes an outcome evaluation infeasible. Too frequently, we set expectations unreasonably high both for demonstration programs and for evaluations. Trying to "overreach" can impede rather than assist the process of social change.

From a foundation's perspective, it is a better long-term strategy to focus scarce evaluation resources on initiatives that can be evaluated (as opposed to all initiatives) and to think about evaluations as a series of demonstration stages that start with "exploratory" demonstrations and move over time to "hypothesis-testing" evaluations.

At our foundation, only a small proportion of what we call "demonstrations" tests hypotheses related to how programs affect specific outcomes. Some of our demonstrations are designed to bring attention to a problem, others are designed to increase confidence that a strategy can be implemented, and still other activities are geared toward assisting the social change process itself.

*Lesson 3: Have a sense of the "evidence standards" for social decisions.*

Not all social decisions require the rigorous inferences that come from social experiments or even from well-designed comparison studies. In recent years, it has become clear that any public decisions involving investments in welfare clients required quite rigorous evidence that the investments would achieve the goals sought. Many recent health care investments, however, have not been subject to such strict evidence requirements. And, in the case of social decisions that get to core ideological beliefs, emotion is often more central than evidence.

In our experience, rigorous evaluations are generally supported and sustained only when the social decision process makes clear demands for rigorous evidence. Evaluation energy and dollars could be conserved if we had a clear sense of the social decision requirements before beginning an evaluation. If a rigorous outcome evaluation, for example, had determined that school-based health clinics reduced the prevalence of high-risk behaviors, this evidence might have persuaded skeptics to adopt a clinic in their schools. For many, however, clinics will never be acceptable because such people base their beliefs on ideological, religious, or moral grounds, not evaluation outcomes. Before investing in a stringent evaluation, researchers would find it useful to know how the results will influence the policy or program process.

*Lesson 4: Analysis from weakly designed evaluations may do more harm than good.*

Controversial programs often allow an opportunity for the inappropriate use of findings. If an evaluation design is fraught with methodological weaknesses, opponents of the intervention will use these shortcomings to discredit the results. Given the possibility of abuse, evaluators should attempt to forestall those situations. There always is room for "exploratory" analysis, but it should be done in times and places that ensure an interpretation as exploratory.

This issue, which requires more attention in the evaluation community, is one of many "fine lines." Except in cases where data come from random assignments or from very careful comparison studies, the chance that selection bias misleads attempts at statistical inference is always significant. Researchers need to be able to distinguish between good studies (even if not perfect) and studies that are likely to do as much harm as good because a weak design leads to questionable inferences. We do not yet have clear guidelines on this distinction.

## References

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally & Company.

Council on Scientific Affairs, American Medical Association. (1993). Confidential health services for adolescents. *Journal of the American Medical Association, 269*, 1420–1424.

Devaney, B., & Rossi, P. (1997). Thinking through evaluation design options. *Children and Youth Services Review, 19*, 587-606.

Dynarski, M. (1997). Trade-offs in designing a social program experiment. *Children and Youth Services Review, 19*, 525-540.

Isaacs, S., & Knickman, J. (1997). *To improve health and health care 1997: The Robert Wood Johnson Foundation anthology*. San Francisco: Jossey-Bass, Inc.

Kisker, E.E., & Brown, R.S. (1996, May). Do school-based health centers improve adolescents' access to health care, health status, and risk-taking behavior? *Journal of Adolescent Health, 18*(5), 335–343.

Kisker, E.E., & Brown, R.S. (1997), Nonexperimental designs and program evaluation. *Children and Youth Services Review, 19*, 541-566.

Lear, J.G., et al. (1991). Reorganizing health care for adolescents: The experiences of the school-based adolescent health care program. *Journal of Adolescent Health, 12*(6), 450–458.

Marks E.L., & Marzke, C.H. (1993). *Healthy caring: A process evaluation of the Robert Wood Johnson Foundation's School-based Adolescent Health Care Program*. Princeton, NJ: Mathtech, Inc.

Metcalf, C.E. (1997), The advantages of experimental designs for evaluating sex education programs. *Children and Youth Services Review, 19*, 507-523.

Miller, B.C., Card, J.J., Paikoff, R.L., & Peterson, J.L. (Eds.) (1992). *Preventing adolescent pregnancy*. Newbury Park, CA: Sage Publications.

Robert Wood Johnson Foundation. (1997). *Annual report, 1996*. Princeton, NJ: Robert Wood Johnson Foundation.

School-based Adolescent Health Care Program. (1994). *The answer is at school: Bringing health care to our students*. Washington, D.C.: School Based Adolescent Health Care Program.

Sonenstein, F.L. (1997). Using self-reports to measure program impact. *Children and Youth Services Review, 19*, 567-585.