

Making Multiple Methods in Evaluation Work for Catholic Hospitals

Four Specific Approaches to Mix Methods Are Available



RYAN EHRENSBERGER, Ph.D. & BY JEFFREY P. MAYER, Ph.D.

From top, Dr. Ehrensberger is administrative director, community benefit, Bon Secours Health System, Richmond, Va.; and Dr. Mayer is assistant professor, Department of Community Health, School of Public Health, Saint Louis University, St. Louis.

Evaluation, whether of process, outcomes or impact, should be an essential component of community benefit programming. It may be used to determine if a program has been implemented with integrity or if the ultimate results met the original goals. However, many community benefit activities are conducted without measurable goals being set and with no evaluation whatsoever being conducted.

One of the challenges that deter evaluation is that the gold standard of evaluation design, where participants are randomly assigned to intervention and comparison groups, is extremely hard for the typical hospital to conduct. Methodological and ethical issues arise. For example, with full-coverage, population-level programs, locating a comparison group may be impossible. Moreover, even when attempted, randomized designs may encounter significant challenges, such as incomplete or inconsistent program implementation, unintended dissemination of the program to the comparison group, high attrition of participants from follow-up measures, and flawed execution of the randomization process. Hence, in all likelihood, quasi-experimental evaluation designs that do not involve randomization will become the standard approach within community benefit.

How can evaluations of community benefit programs remain credible if the randomized gold standard design is not feasible? One answer is to triangulate information about a program using multiple methods and multiple data sources. By piecing together information gathered using dif-

ferent methods, the limitations or biases of any one method can be overcome. Ideally, results from different methods converge on a single conclusion concerning a program's effect. For example, qualitative data from focus groups may complement results from a pretest-posttest phone interview, or data from administrative records may reinforce findings obtained from a self-report survey of program participants.

This article introduces some basic approaches to triangulation — evidence from more than one source — in program evaluation, provides some examples of the use of multiple methods in community benefit evaluations, and discusses some of the strengths and limitations of a multiple method approach. Four specific ways to mix methods will be introduced:

- 1) combining qualitative and quantitative methods
- 2) using complementary quantitative data
- 3) deploying “patched-up” or hybrid designs
- 4) employing program theory

COMBINING QUALITATIVE AND QUANTITATIVE METHODS

Qualitative data addresses evaluation questions concerning how and why program effects are achieved. By doing so, it may complement quantitative data, which addresses questions concerning the size of a program's effects and whether these effects are greater than chance. Focus groups or in-depth interviews conducted with participants after a program can illuminate the sequence of social, cognitive, environmental and behavioral changes brought about. For example, in a nutrition education program, qualitative data from participants described how they actively learned how to shop for healthy foods and came to understand the relationship between poor eat-

ing and chronic disease. They also built support for healthy eating among their family and friends. The program may also enhance confidence in other results showing an improvement in healthy eating behavior at a six-month post-test. Other useful applications of qualitative approaches in program evaluation include identifying unanticipated outcomes, recognizing reinventions and/or adaptations of the program model, and uncovering implicit program theory.

In a community benefit program evaluation at Bon Secours Richmond Health System in Richmond, Va., Ryan Ehrensberger (co-author of this article) used qualitative data to bolster conclusions about the effectiveness of a health care provider quality improvement program.¹ Results from a pre/post chart review showed moderate to large improvements on seven key outcomes, but limitations to the evaluation included lack of a comparison group and small sample size. However, during in-depth qualitative interviews, providers reported fully engaging with the program's various components and making substantial progress in successfully managing their patients' disease as a result of the quality improvement initiative. Also, providers reported they were not significantly exposed to other asthma management messages or activities external to the quality improvement intervention, strengthening the argument that this program was responsible for the observed outcomes.

Given these findings, Ehrensberger and Bon Secours' colleagues plan to continue the program and to remedy some of the shortcomings in the design of the initial pilot evaluation.

USING COMPLEMENTARY QUANTITATIVE DATA

As previously noted, including a comparison group when evaluating community benefit programs is often not possible. Withholding the intervention may be unethical, or it may be difficult to locate a comparable population in terms of both context and composition. In these situations, where a pre/post design would seem the sole option, history (or secular trend) looms as a major threat to internal validity. That is, without a comparison group, it would be difficult to determine if other public health initiatives, messages in the media, or other factors, are responsible for any pre/post change rather than the intervention program of interest.

One approach to this problem is to gather and analyze complementary data concerning local media's coverage of health topics to characterize secular trend and judge its influence. In effect, trends in health reporting by the media serve as proxy measures of this key internal validity threat.

How can evaluations of community benefit programs remain credible if the randomized gold standard design is not feasible? One answer is to triangulate information about a program using multiple methods and multiple data sources. By piecing together information gathered using different methods, the limitations or biases of any one method can be overcome.

In an example of this approach, Jeffrey Mayer (co-author of this article) evaluated a campaign to improve childhood immunization in a medically underserved rural community.² Data on receipt of eight childhood vaccines were obtained before and after the campaign from preschools, local health departments and medical practices. Following adjustment for birth order and demographics, at post-intervention a significantly greater proportion of children were up to date for six of the eight vaccines. Coverage of child health topics in 23 small-town newspapers was tracked during a concurrent 30-month period. Overall, press coverage of child health was low, with only 3 percent of all health-related articles focusing on this topic. Time-series analyses revealed an absence of trend in the newspaper data, indicating that reporting on child health did not increase during the course of the campaign. With no trend indicated, it becomes less likely that history provides a plausible alternative explanation to the campaign's effects.

Of course, other ways to supplement an evaluation with additional quantitative data are possible. Archival records may reinforce findings from patient and provider surveys, or a non-equivalent dependent variable may be employed in place of a comparison group.

DEPLOYING "PATCHED-UP" OR HYBRID DESIGNS

Frequently, evaluators restrict themselves to a single design, typically choosing one of the classic design options offered by Campbell and Stanley.³ However, it is possible to combine more than one design in a single evaluation project. Application of multiple designs is sometimes known as "patched-up" or hybrid design. In this approach, it is hoped that the weaknesses of one design will be offset by the strengths of another. If results

Application of multiple designs is sometimes known as “patched-up” or hybrid design. In this approach, it is hoped that the weaknesses of one design will be offset by the strengths of another. If results from the designs converge, then the case that the program caused any observed improvement in outcomes is strengthened.

from the designs converge, then the case that the program caused any observed improvement in outcomes is strengthened.

In the simple example shown below, X represents an intervention to increase physical activity among eligible obese diabetic patients, and O represents a self-report diary-based assessment of leisure-time physical activity. The intervention is first implemented at Clinic A with only a posttest. At Clinic B, the self-report diary measure is administered both before and after the program, with the pretest corresponding temporally with Clinic A’s posttest.

CLINIC A: X O₁ [Design 1]

CLINIC B: O₂ X O₃ [Design 2]

Clearly, history (or secular trend) is a major threat to validity for both designs. Even so, if the intervention had an impact, O₁ should be greater than O₂, and O₃ greater than O₂. Since O₁ and O₂ are measured at the same time, the history threat is less plausible. That is, O₁ includes the effects of both history and the intervention, and O₂ just history. Patched-up or hybrid designs often fit well with the situations likely to be encountered by community benefit programs. Here, for example, is a situation where programs start small and scale up to additional settings over time. In addition, denying the intervention to patients is avoided.

EMPLOYING PROGRAM THEORY

Program theory provides evaluators and other stakeholders with a “plausible and sensible model of how a program is supposed to work.”⁴ It details the sequence of short, intermediate and long-term effects of the program, and identifies the paths and causal chains to be activated. Often, this sequence is depicted in a logic model.

Theory-based evaluation augments traditional evaluation by introducing multiple methods that seek to confirm whether the anticipated sequence of program effects has actually been set in motion.

Evidence that a program’s theory has been activated can enhance a quasi-experimental evaluation’s credibility. For example, Ricardo Wray and colleagues, in a single-group, posttest-only evaluation of a radio campaign to increase walking in a Missouri town, demonstrated that greater exposure to the campaign was associated with improvement in knowledge and beliefs concerning the social and health benefits of walking, which, in turn, was associated with greater levels of walking behavior.⁵ If the study had only included data on the walking outcome, the internal mechanism of the program would have remained unexplored, and the assertion that the program produced an increase in walking would be less convincing.

Another significant benefit of theory-based evaluation is that it allows for identifying which causal chains worked and which did not. This permits redevelopment of programs by dropping poorly performing chains and retaining successful ones. In this way, program development can become an ongoing process. In response to initial evaluation results, some chains are dropped and others added, with subsequent evaluation findings potentially leading to even further program fine-tuning.

Finally, program theory can facilitate evaluation planning. By selecting specific indicators (or measures) for each anticipated short, intermediate and long-term outcome, an initial blueprint for evaluation is produced. With a thoughtful and carefully constructed program theory, debates about what to measure and what not to measure are avoided because each selected measure clearly taps an indispensable aspect of the program’s overall conception. A lack of program theory often signals a lack of shared understanding among stakeholders about how a program is supposed to work, and often results in setting unrealistic objectives and goals for a program.

CONCLUSION

This article has introduced some possible ways to make use of multiple methods in program evaluation. Many other possibilities and examples are available.⁶

The purpose of triangulation is either congruence or complementarity.⁷ For congruence, the aim is to obtain similar results with each method. The two-clinic “patch-up” evaluation design discussed above is an example of congruence. For

complementarity, results from one method enrich, expand or clarify results from another. Here, one method is primary and the other secondary. Ehrensberger's use of in-depth interviews to reinforce findings from a primarily pre-post chart review evaluation is an example of complementarity. Both congruence and complementarity can have an important role in improving the credibility of quasi-experiments.

However, there are challenges to making full use of multiple methods. First, results from different methods may not agree. Obviously, this may produce opposing views concerning the desirability and effectiveness of a specific program, and may complicate decision-making. Second, results from multiple methods may converge misleadingly if the two methods share the same sources of error. Third, multiple methods can be time-consuming and expensive. With program theory approaches in particular, where data on a potentially large set of intermediate outcomes is required, the burden can be especially large. But in many cases, a small investment of resources can help remove uncertainty about an evaluation's findings.

Although Catholic hospitals have a long and rich tradition of community service, evaluation of community benefit is a relatively new undertaking. A case study of several community benefit programs in California and Texas in 2006 noted "lack of familiarity with evaluation methodologies and the reality of the substantial challenges in establishing measurable results."⁸ To begin to raise program evaluation skill and capacity, CHA is currently piloting the "Evaluation Guide for Community Benefit Programs," a new resource for its members. The guide introduces key evaluation concepts and provides a basic step-by-step process. CHA has also established a work group on assessing program effectiveness to help with the pilot of the guide and to advise CHA on program evaluation more generally. The ideas introduced in this article are intended to complement the material in the guide and support the efforts of the work group.

For the emerging field of community benefit program evaluation, multiple methods represent an exciting opportunity for creative and innovative application of evaluation techniques to solve common challenges in quasi-experimentation. ■



**Comment on this article
at www.chausa.org/hp.**

NOTES

1. Ryan Ehrensberger, "Evaluating the Effectiveness of Community Benefit Programs Case Study: Controlling Asthma in the Richmond Metropolitan Area" (presented at the Association for Community Health Improvement Conference, Atlanta, 2008).
2. Jeffrey Mayer, "Using Media Data to Interpret Secular Trend as a Threat to Internal Validity: An Application in Community Health" (presented at the 6th International Union for Health Promotion and Education Conference on the Effectiveness and Quality of Health Promotion, Stockholm, Sweden, 2005).
3. Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963).
4. Leonard Bickman, *The Functions of Program Theory: Using Program Theory in Evaluation: New Directions for Program Evaluation* (San Francisco: Jossey-Bass, 1987), 5-27.
5. Ricardo Wray, Keri Jupka and Cathy Ludwig-Bell, "A Community-Wide Media Campaign to Promote Walking in a Missouri Town," *Preventing Chronic Disease*, www.cdc.gov/pcd/issues/2005/oct/05_0010.htm.
6. Melvin M. Mark and R. Lance Shotland, *Multiple Methods in Program Evaluation: New Directions for Program Evaluation* (San Francisco: Jossey-Bass, 1987); Jennifer Greene and Valerie Caracelli, *Advances in Mixed Method Evaluation: New Directions for Program Evaluation* (San Francisco: Jossey-Bass, 1997); and John Creswell and Vicki Plano Clark, *Designing and Conducting Mixed Methods Research* (Thousand Oaks: Sage Publications, 2007).
7. Jennifer Greene and Charles McClintock, "Triangulation in Evaluation: Design and Analysis Issues," *Evaluation Review* 9 (1985): 523-545.
8. Arthur Himmelman, *Advancing the State of the Art in Community Benefit (ASACB): Phase Two Final Evaluation Report* (Minneapolis: Himmelman Consulting, 2006), 16.