POLICY STUDIES ASSOCIATES, INC.
1718 CONNECTICUT AVENUE, N.W. • SUITE 400 • WASHINGTON, D.C. 20009
(202) 939-9780 • (202) 939-5732 FAX • WWW.POLICYSTUDIES.COM

# Using Evaluation Methods to Promote Continuous Improvement and Accountability in After-School Programs: A Guide

Elizabeth R. Reisner

As demand for after-school services has grown, the public has responded with support for growth in federal, state, and local funding for these services. As in other successful grassroots movements, this high level of support is evident across groups with diverse objectives. Some after-school advocates are mainly interested in increasing the supply of safe, dependable after-school environments for the children of working parents. Others care most about using after-school programming to give children and youth positive developmental experiences that extend beyond academic learning. A third group is most concerned about expanding the learning opportunities available to low-achieving students and, in some instances, in using the after-school hours for remediation that will improve students' measured levels of achievement.

Although these diverse interests have joined forces around their common objective of increasing and improving after-school opportunities, they tend to hold different expectations for programming and for the results that after-school programs might achieve. The experiences of successful after-school initiatives prove that multiple interests can be successfully accommodated, however, and, despite their diversity, can enrich program development and implementation. Even so, a multiplicity of stakeholder voices can present challenges to after-school program operators.

One challenge arises in the evaluation of these programs. However much diverse stakeholders may find common ground in the design of their after-school program, they will almost certainly seek different types of information to determine whether the program is achieving the purpose most important to them. To address these different perspectives, the evaluator must listen to and distill diverse priorities and then develop data collection, analysis, and reporting strategies that will inform all major stakeholder groups, while providing the program with directions for improvement.

Knowledge of a few key steps can help program operators and evaluators design and conduct evaluations that address the major interests of program stakeholders. This paper describes a series of steps for designing and conducting such evaluations, based on the author's experience in evaluating after-school programs and other education and youth-development initiatives. The paper provides examples of these steps from a current evaluation of a large local after-school program, that of The After-School Corporation (TASC) in New York. Interim findings from this evaluation are available in Reisner, Russell, Welsh, Birmingham, and White (2002) and Welsh, Russell, Williams, Reisner, and White (2002); the final report of this evaluation will be issued later in 2004.

## Setting the Parameters for Evaluation

**Develop a program theory or logic model that describes the initiative's goals, its strategies for achieving its goals, its expected outcomes, and the anticipated timetable for program implementation and results.**

Program theory is important because it defines the resources, interventions, and conditions necessary for the program to achieve its intended results. These elements tell the evaluator, first, what program intervention is being evaluated and what its key components are, and, second, what information to collect in order to evaluate whether the program operates as planned. Weiss (1995) summarizes the reasons for basing evaluation in solid theory:

> The concept of grounding evaluation in theories of change takes for granted that social programs are based on explicit or implicit theories about how and why the program will work. . . . The evaluation should surface those theories and lay them out in as fine detail as possible, identifying all the assumptions and sub-assumptions built into the program. The evaluators then construct methods for data collection and analysis to track the unfolding of the assumptions. The aim is to examine the extent to which program theories hold. The evaluation should show which of the assumptions underlying the program break down, where they break down, and which of several theories underlying the program are best supported by the evidence.

Basing evaluation in strong theory also increases the likelihood that evaluation findings will contribute to sound decisions about program management and operations.

For after-school programs that aim to produce concrete benefits for youth, an important element of the change theory is the estimation of the magnitude of benefit that student participation in the after-school services might plausibly stimulate. With this information, the evaluator can make design decisions that will permit the evaluation to detect student-level changes of the estimated magnitude. As Kane (2004) has shown, after-school evaluations that measure achievement gain have not typically been capable of detecting the small improvements in achievement that might plausibly be attained by programs that only fractionally increase academic learning opportunities for youth.

Also, although many after-school programs aim to provide safe, supervised care for students during hours in which they might otherwise be on their own, relatively few after-school program evaluations actually address this program goal. Integrating the goal into the program's change theory is the first step in ensuring that the evaluations measure change in student exposure to safe, supervised care after school.

**Bring stakeholders into the evaluation process by taking account of their shared goals and concerns and by helping them meet their own information needs. Make sure that stakeholders see how evaluation will help them do their jobs better.**

As already suggested, a special challenge in developing a program theory is that stakeholders' views on the program's underlying goals and logic may differ significantly. Sometimes the process of delineating a program theory uncovers a multiplicity of perceptions about the after-school program and the purposes of the evaluation. Evaluation planning should not move forward until those differences are acknowledged and essentially resolved.

To make the evaluation useful to various stakeholders, the TASC evaluation is spinning off tailored evaluation reports to key stakeholder groups. These reports synthesize evaluation data and findings that are of special relevance to particular groups. So far, the evaluation has published short reports on TASC's student participants, school principals, TASC staff, parents of TASC participants, and TASC programs that receive support under the federal 21st Century Community Learning Centers program. Upcoming special reports will focus on TASC programs in middle and high schools. Although valuable, the preparation of separate reports for different audiences adds to evaluation costs.

## Deciding What to Evaluate

**Using the program theory, pinpoint the events, activities, relationships, and outcomes that are central to the initiative's success. Focus data collection in these key areas. Don't try to collect data on everything that is "interesting." Use the program theory to determine when to expect both immediate and long-term effects.**

A benefit of theory-based evaluation is that it helps identify the relationships in the program model that are likely to be the weakest, thus permitting evaluators and others to focus evaluation resources on the program's activities, capacities, and outcomes that are particularly problematic. The program theory should also indicate the amounts of program participation believed essential to produce the expected results. This information tells the evaluator when and under what conditions it is reasonable to look for the intended results.

Programs that aim to promote learning growth may want to measure important interim indicators of learning, such as gains in school attendance, homework completion, parent involvement, and other early signals of students' engagement in learning (Kane, 2004). Changes in these areas are likely to emerge before learning effects show up on test scores, and they may be easier to measure in communities where comparable annual achievement data are not available at the targeted grade levels.

An evaluation will be more effective if it conforms to the timing of change patterns predicted by the program's change theory. The TASC evaluation emphasized project start-up and implementation in its first phase, project-level service quality in its second phase, and initial student-level changes in its third phase. This explicit schedule concentrated evaluation resources in areas where change was anticipated, avoiding the needless direction of resources to the measurement of outcomes that the change theory did not anticipate would occur until later.

Without determining what duration of program participation is necessary for results, an evaluation can find its reporting schedule trapped by the demands of an external policy-setting body, such as a school board, a state legislature, or a federal agency. Federal programs, for example, are often expected to show student progress in time to inform program reauthorizations, even though reauthorization schedules are typically set without regard to the requirements of real-life implementation.

**Consider the program context.**

After-school programming adds to and can enrich the everyday instructional activities of schools. A fact of life in today's schools is the high stakes attached to student performance on standardized tests of academic achievement. For an after-school program to affect student achievement on tests, course grades, and other measures, it must link to, extend, and reinforce the learning sought during the regular school day. After-school experiences that are not tied to regular school-day instruction may enrich a child's cognitive development but may not affect conventional measures of school achievement. This doesn't necessarily mean that program designs should be changed to emphasize the skill development that is measured on tests. It does mean that goals and activity priorities should be set with a clear understanding of the child or youth outcomes they are likely to affect, including whether these are measurable using standardized achievement tests.

**Consider the program's intended effects on socialization and resilience. If a central goal of the after-school program is to strengthen students' capacities in areas such as developing relationships with peers, planning a successful future, and avoiding risk behaviors, build in measures of student growth in these areas.**

Researchers (including those at the Developmental Studies Center, Public/Private Ventures, the University of Wisconsin Center for Education Research, and PSA) have developed and shared easy-to-administer measures of positive youth development, typically in the form of survey scales, and have published the properties of these measures. Any evaluator can use these measures now, with some certainty that they assess the psychosocial domains that they are intended to capture.

It is not unusual for a program theory to postulate that participants' growth in areas such as socialization and resilience will precede growth in areas such as academic achievement and the capacity for and interest in planning for the future. If so, it may not be reasonable to look for change in longer-term outcomes without first seeing evidence of change in student attitudes and relationships. Where this is the case, evaluation timetables and measures need to consider these anticipated sequences of change.

## Crafting the Evaluation Design

**Attend to research-design principles that ensure accuracy, generalizability, and lack of bias.**

Important research-design principles include basics such as:

- Consulting prior research and adopting methods and instruments that have proven successful in similar situations

- Pilot-testing data collection instruments to make sure that those who must provide information understand them and that all of the intended respondents are likely to understand them in the same way

- Establishing sample sizes that will achieve the levels of precision and statistical confidence needed to answer the study's research questions

- Minimizing the evaluation's burden on respondents by querying only those respondents who are well positioned to answer honestly and accurately and by using administrative data and other existing information whenever possible, so long as the data are reliable and comprehensive

- Designing data collection to achieve high rates of response from intended respondents; designs that generate high response rates permit smaller sample sizes if the samples are representative of the populations being measured

- If possible, employing designs that permit the attribution of causality (by permitting all other explanations for program-related change to be considered and ruled out)

The TASC evaluation illustrates deliberate choices made in addressing each of these principles. For example, one choice centered on the need to obtain a high response rate on the annual survey of TASC site-level project directors (who are known in the program as site coordinators). In developing the evaluation design, it became clear that site coordinators were the best sources for annual data on many aspects of project activities, relationships with the host schools, relationships with parents, project staffing, and other topics. However, asking site coordinators all of the questions for which the evaluation needed answers would produce a long, burdensome survey that would itself reduce the likelihood of obtaining timely responses. The evaluation had to balance its need for accurate, comprehensive information with its need for very high survey response rates. The evaluation took three steps to reach the right balance, it (1) gave site coordinators the option of completing the survey either online (with many features built in for ease of on-line survey completion) or in traditional paper-and-pencil form, (2) scrutinized every survey item, deleting those that might be misunderstood or might not produce policy-relevant data, and (3) communicated to site coordinators the reasons that their survey responses were important. These efforts resulted in a very high response rate from site coordinators.

## Establish frameworks for accurate comparison.

Evaluators can select from among many good approaches for comparing outcomes. All of these approaches share an important trait, which is that they allow an evaluation to demonstrate the difference between outcomes likely to occur with and without a given program intervention. This information is essential to demonstrate the changes associated with implementing an after-school program. But ascribing any observed change in performance to the after-school program requires careful attention to the selection of a comparison group.

In particular, the evaluator must be able to rule out alternative explanations for differences emerging between participants and nonparticipants. This means that participants and nonparticipants must be as similar as possible in every way except exposure to the program. In addition to making sure that participants actually participate in the program and that nonparticipants do not, the evaluator must be certain that the two groups are not different in other ways, such as their pre-existing level of academic achievement, personal motivation, family background, and so on.

The most rigorous method for creating identical participant and nonparticipant groups is to use random assignment to determine which students will enroll in the after-school program and which will not. Implemented carefully, this procedure produces a subject distribution in which the factors associated with desired program outcomes can be assumed to be randomly and equally distributed across the participant and nonparticipant groups.

Evaluators and program managers sometimes run into problems in implementing a random assignment design, however. Program administrators may look askance at random assignment because it may require denying after-school services to children and youth who want to participate. Another common problem emerges when virtually all students in the intended target group already participate in services. Random assignment in these instances means that some students will be denied services while there are empty service slots or that members of the nonparticipant control group may on their own find the after-school service they are seeking from another provider that is not part of the evaluation.

In response to problems such as these, evaluators sometimes use other research methods as alternatives to randomization. The methods used most frequently are:

- Comparing matched pairs, in which the evaluator matches each participant in the evaluation sample with a nonparticipating youth who is as similar as possible to the participant on all indicators associated with program outcomes

- Making adjustments to participant and comparison groups to offset prior differences on attributes believed to be associated with the desired program outcomes

The preceding methods cannot, however, address unmeasured differences among students, especially in areas such as motivation and family support, as effectively as can random assignment.

## Planning for Data Collection

**Identify and follow all school district and youth service agency requirements for the protection of human subjects in the handling and reporting of data. Secure parental permission before obtaining personally identifiable data about students. Maintain the confidentiality and anonymity of these data.**

Parents need to be willing for questions to be asked of their children and also for their children's school records to be reviewed. In addition, parental permission is required under various provisions of law and local policy. Parents are most likely to give permission if they understand the purpose of the research and the limitations on the use and distribution of the data. Requesting parental permission at program registration gives program operators and evaluators a good opportunity to answer parents' questions about the research and about any effects it might have on their children and the program more generally.

If the after-school program is school-based, it will also be important to obtain approval from the district research office or institutional review board. Typically, the school district will want to be consulted if data are to be collected from students, parents, or district employees, and if any data collection activities are conducted during school hours. Arrangements with the district research office will be necessary to obtain information from the school district's student information systems, such as school attendance and test scores.

**Obtain information from the most reliable sources available, consistent with reasonable evaluation budgets.**

If an evaluation needs student-level data in areas such as attendance, test scores, and grade promotion, administrative records from a central source are likely to be more accurate and extensive than teacher- and school-level records or personal recollections by teachers, parents, or students. Moreover, use of centralized sources will generally minimize burden on persons involved in the program.

For consistent, comparable data about activities and program contexts, the best sources are researchers' systematic, structured observations. However, the cost of training observers and of repeating observations and ratings may preclude all but the most selective use of this data-collection method. Surveys of key informants (such as project directors and student participants) may generate data that are somewhat more self-interested than observation data but much less expensive to collect, code, and analyze.

### Describe the children and youth who participate in the after-school initiative.

After-school programs typically set out to serve particular populations of students, either students who attend a particular school, students who live in a particular neighborhood, or students with some other set of shared characteristics (e.g., an interest in sports or the arts). A program theory of change is typically premised on an intent to serve a particular population and to address their shared needs or interests. Too often, however, no effort is made to determine the characteristics of the actual children and youth who enroll and participate in the after-school program. The program simply assumes that it is serving its intended target group.

Understanding the characteristics of the students served in an after-school program can inform an evaluation in useful ways. For example, information on students' family income (as measured most frequently by eligibility for free or reduced price lunch) and on their prior academic achievement can indicate the overall level of disadvantagement of the served population. The prior achievement measure can also serve as an essential baseline if the evaluation measures achievement change. In addition, information on student demographic characteristics is important program feedback on its own. It tells the program whether it is serving the types of students it originally set out to serve.

The TASC evaluation has provided feedback to TASC on student characteristics from the inception of the evaluation. Because the evaluation uses the New York City Department of Education's database to track student characteristics, it can describe the demographic and educational-performance characteristics of both participating students and nonparticipants who are enrolled in the schools that host the TASC programs. Student characteristics available through that system and used in the TASC evaluation include students' gender, race/ethnicity, eligibility for special education services, eligibility for free or reduced-price lunch, grade in school, school attendance rates, reading and math scores on citywide tests, and whether the student is a recent immigrant.

**Gather data on student exposure to the after-school program and use the data in analyzing program implementation and effects.**

Because students are not typically required to participate in after-school activities, their after-school attendance is likely to be more sporadic than their regular school attendance. In assessing program effects, it is important to consider the extent to which effects are linked to students' level of program participation (or "dosage"), since the program theory is likely to link magnitude of student outcomes to the amount of after-school participation.

The TASC evaluation uses student identification numbers assigned by the New York City Department of Education to link student data from the TASC after-school enrollment and attendance system to information from the Department's student data system. In effect, the evaluation brings the two data systems together using school and student identification numbers as the linking device. With this merged data set, the evaluation has charted the characteristics and educational progress of after-school participants based on the frequency and duration of their after-school participation. Analyses of these data have shown greater effects for participants with high levels of after-school attendance, when these students are compared to similar students with less participation.

**Set up a system for managing information. Organize the system to assemble data at the level that is most relevant to the initiative (e.g., child, school/provider, community).**

A common problem encountered by inexperienced evaluators is needing to organize and analyze a large volume of data but having no integrated system in place for data management. Although many software options are available for managing evaluation data, they all rely on careful anticipation of key research questions and reporting needs. At the same time, the data management system should be flexible, in order for the evaluator to pursue and report on unanticipated issues or findings as the evaluation proceeds.

## Planning for Analysis and Reporting

**Provide feedback from the evaluation to program operators and stakeholders early and often. Link reporting schedules to the timing of stakeholders' needs for information. Focus analysis and reporting on elements that program operators can change, not on conditions that are beyond the capacity of the program to affect.**

The regular transfer of evaluation findings to program stakeholders is an important part of what defines "continuous-improvement" evaluation. Gray (1993) describes the key purpose and methods of such evaluation:

> Evaluation is an important ongoing process that supports the organization striving for excellence in the achievement of its mission. It is a process of asking good questions, gathering information to answer them, and making decisions based on those answers.

An important element in making evaluation findings useful is generating and reporting them on schedules that correspond to important decision points, such as annual contract-renewal and hiring periods. TASC has used evaluation information on the importance of staff satisfaction and sense of professionalism to inform decisions about staff training, hiring and retention, and working conditions. Because early findings from the TASC evaluation suggested that high levels of staff satisfaction were associated with high levels of student engagement and enjoyment of the after-school experience, TASC took steps to raise staff satisfaction by improving the quality of staff training and encouraging positive efforts by the program's grantee organizations on behalf of project staff.

**Use the evaluation as an opportunity to analyze and report on promising local practices. Identify reporting topics based on areas in which program operators are searching for help.**

Given the lack of good information on effective practices in after-school programming, evaluations of these programs should, when feasible, generate information on the outcomes associated with specific practices. But because such information is generally not available until several years after an evaluation has begun, stakeholders are likely to need preliminary information sooner. One approach is for evaluators in early evaluation phases to identify and describe practices that meet criteria of relevance (as measured by whether the practice addresses a high-priority need) and quality (as measured by whether the practice is consistent with other research on positive child or youth development).

In the second and third years of the TASC evaluation, the evaluation team developed and published a series of "resource briefs" on promising practices observed in TASC projects. The briefs each addressed a high priority need, such as recruiting and training qualified after-school staff, improving the quality of after-school homework assistance, and using project-based learning to build student engagement and skill development. The briefs (each 4 to 7 pages in length) reflected early findings from the evaluation but were written in a journalistic style, drawing quotes and examples from after-school staff in TASC projects known to be making serious efforts to address the topic area intelligently. TASC has disseminated the briefs widely in technical assistance and training sessions. In its final year, the TASC evaluation is revisiting the identification of promising practices using empirical means.

**Make sure that program reports convey information in language that is readily understandable to key stakeholders.**

If an evaluation is to be truly useful to several audiences, the evaluator may need to prepare several different types of reports. These are likely to include one or more complete technical reports as well as short, non-technical summaries and possibly even shorter summaries on topics of special interest. In spring 2002, when TASC was engaged in a struggle to restore city funding for after-school services, the evaluation developed several "information briefs" that described how (1) the program was indeed delivering the types of services that it claimed to be providing, (2) large numbers of students were participating in these services, and (3) the program had made a serious commitment to public accountability through its external evaluation. Not surprisingly, the summaries sparked deeper questions that led policymakers to analyses and findings in the evaluation's technical reports.

## **References**

Gray, S.T. (1993). "A vision of evaluation." In S.T. Gray (Ed.), *Leadership is: A vision of evaluation*. Washington, DC: Independent Sector.

Kane, T.J. (2004). *The impact of after-school programs: Interpreting the results of four recent evaluations*. A Working Paper of the William T. Grant Foundation.

Reisner, E.R., Russell, C.A., Welsh, M.E., Birmingham, J., & White, R.N. (2002). *Supporting quality and scale in after-school services to urban youth: Evaluation of program implementation and student engagement in the TASC after-school program's third year*. Washington, DC: Policy Studies Associates.

Weiss, C.H. (1995). "Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families." In J.P. Connnell, A.C. Kubisch, L.B. Schorr, & C.H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts*. Washington, DC: The Aspen Institute, Roundtable on Comprehensive Community Initiatives for Children and Families.

Welsh, M.E., Russell, C.A., Williams, I., Reisner, E.R., & White, R.N. (2002). *Promoting learning and school attendance through after-school programs: Student-level changes in educational performance across TASC's first three years*. Washington, DC: Policy Studies Associates.

January 26, 2004