5

Responding to Accountability Requirements While Promoting Program Improvement

Blanca Flor Guillén-Woods, Monica A. Kaiser, Maura J. Harrington

Abstract

The impact of No Child Left Behind (NCLB) is usually understood in relation to schools and districts, but the legislation has also affected community-based organizations that operate school-linked programs. This case study of an after-school program in California demonstrates how educational accountability systems that emphasize students' academic achievement and scientifically based research prompted evaluators to modify evaluation questions, methods, and analytic techniques. The external demands of NCLB transformed the evaluation to support the relevance and value of this community-based program within the evolving framework of accountability for the school. © Wiley Periodicals, Inc.

Recent emphasis on the No Child Left Behind legislation (NCLB, 2001) and its research guidelines for evaluation has resulted in increased outcome evaluation activities in K–12 settings. Under NCLB, schools have been held accountable for student academic achievement through use of standardized test scores. Though these accountability efforts have not always

shown a strong relationship with school improvement (Nichols, Glass, & Berliner, 2005), they have succeeded in creating a school environment that is increasingly focused on high-stakes testing. In this environment, even community-based organizations with school-linked programs, not directly under NCLB, feel the pressure to demonstrate explicitly their contribution to improved student academic performance. Over the last 5 years, this increased focus on measuring academic progress through standardized tests has had positive and negative impacts on local educational agencies (LEAs) and community-based organizations (CBOs) and their local evaluators.

This chapter explores this impact through a case study of a program in Los Angeles to illustrate how evaluation decisions have been indirectly affected by NCLB and, more specifically, scientifically based research (SBR). The case study presented here is of Woodcraft Rangers (WR), a community-based organization providing after-school services at school sites in southern California.

Brief Overview of Woodcraft Rangers

The primary goal of the Woodcraft Rangers after-school program is to extend schools' capacities to provide a safe, supportive environment beyond the school day and help students improve social, behavioral, and learning skills that contribute to improved school achievement. The Woodcraft program model draws on best practices in the after-school field and on research-based practices identified by organizations such as the National Youth Development Center, Clark and Associates, Public/Private Ventures, the Harvard Family Research Project, and others. Woodcraft Rangers has offered after-school programs in more than 40 schools in Los Angeles County through funding from the California Department of Education's (CDE) After School Education and Safety Program (ASES), U.S. Department of Education's 21st Century Community Learning Center Grants, and Los Angeles City funds. To meet the program's evaluation needs, including their funding agencies' requirements, Lodestar Management/Research was contracted in 1997 and continues to work on the program's current evaluations.

Woodcraft Rangers Case Study

This case study focuses on how WR and its evaluator-modified evaluation activities to measure the program's progress toward program improvement and meet the reporting requirements of the funders. Further, the case study describes how evaluation decisions and approaches were indirectly affected by the NCLB legislation. Although this case study illustrates only specific changes in the evaluation activities of one school-linked program, the evaluators observed similar changes in several school-linked programs.

General Evaluation Approach. During the late 1990s, Woodcraft was in its early stages of evaluation—collecting information on the number of

students who attended the program, their demographics, retention rate and program satisfaction level. By 2001, even prior to the signing of NCLB, the program's leadership was aware of the shifting school environment and began feeling pressure from partner schools for increased evaluation efforts focused on academic outcome data for participants. It was at this point that WR began working most intensively with its evaluators. The first step in designing the evaluation was identifying the research questions. WR had several program goals and key stakeholders. Similar to other programs, WR also had limited funds for evaluation, making it a challenge to answer all of their research questions each year. Thus, the core evaluation questions had to be identified. One of the key questions asked during this phase was, "Does the value of the information outweigh the effort to obtain it?" In the era of NCLB, the answer to this question was a resounding yes when measures of academic performance were included.

The evaluator worked collaboratively with WR to create an evaluation plan that incorporated core research questions related to academic achievement, youth development, and program improvement (see Table 5.1). Note

**Table 5.1. Evaluation Questions Before and After
No Child Left Behind**

Initial Evaluation Questions	*Evaluation Questions Refined Over the Years for Program Improvement*
Are youths and parents satisfied with the program? Did we reach our objectives to: – Increase social and communication skills? – Develop youths' acquisition of new skills and knowledge? – Increase self-confidence?	1. What population is WR serving? 2. To what extent does WR increase youth engagement in after-school activities? 3. Do WR programs help schools keep students safely occupied during after-school hours? 4. To what extent are school administrators, parents, and participating students satisfied with the program quality? 5. Do students who participate in WR attend school more regularly? 6. To what extent do students who attend WR develop prosocial interests and behaviors and avoid at-risk behaviors? *Enhancements to the Evaluation Influenced by NCLB* 7. To what extent do WR participants improve their attitude toward school and learning? 8. To what extent do WR participants improve their learning skills and habits? 9. Is participation in WR associated with a higher level of academic achievement?

that only three questions are directly related to academic performance (Questions 7 to 9); the others are more closely related to the outcome areas WR directly addresses or to their process evaluation questions.

Although NCLB facilitated the program's decision to include standardized test scores, SBR guidelines indicated that the gold standard for evaluation was in using experimental methods, especially random assignment (DOE, 2003). Though not the only option, preference for this approach was clearly recognized by the schools, the community-based organizations providing services to the schools, and, of course, their evaluators. All three entities realized that even if not required, they should employ this methodology whenever possible to be competitive for funds and demonstrate their effectiveness.

WR was no exception, as the program decided to include a comparison group design to address some of their research questions. Despite the emphasis on random assignment, WR would not deny program access for the purposes of evaluation. To navigate the challenge of creating a comparison group, the evaluators decided to use a post-hoc comparison group at the end of each program year by means of district databases. Agreements with the district for data sharing on participants extended to data sharing on all students in a Woodcraft school. In fact, one of the unexpected benefits of NCLB was the opportunity for community-based programs and districts to collaborate more closely in accessing academic data electronically. That is, as NCLB mandated that school districts improve their systems for tracking student achievement, much more data became accessible to CBOs with school-linked programs. Using these data, analyses comparing participants to nonparticipants (statistically controlling for demographic and baseline measures) became the centerpiece of the WR evaluation following NCLB. Over the years, these analyses positively affected the program because WR was able to demonstrate that it had a legitimate, important role on school campuses. The academic findings were particularly useful because they were presented along with results from the paired sample *t*-tests examining changes on participants' social skills, at-risk behavior, and attitude toward school.

The local evaluator and WR collaborated to conduct this type of evaluation for several years while, at the same time, addressing the challenge of including process and other outcome measures with essentially the same evaluation budget. Although academic measures were furnished electronically by the school district for participating schools, these data still had to be merged across years and matched with WR records for participating students. Completing this process for more than 30 schools, as well as recoding data, computing changes, and selecting comparison groups, was extremely time-intensive. Although the program wanted to continue including a comparison group design given the weight it held for WR stakeholders, the cumulative evidence from these analyses suggested that compared to nonparticipants, participants made significant academic gains

only after they were in the program for 6 months or more (Lodestar, 2005). Thus, the program was interested in finding ways of attracting more students and retaining them for 6 months or more. WR became interested in additional process-related questions: (1) What attracts students to WR? (2) What individual-level characteristics are related to how long students attend WR? and (3) What program-level characteristics are related to how long students attend WR? Because WR's evaluation budget was limited, it was not possible to conduct the full-impact evaluation while conducting an in-depth process evaluation. The solution reached was to focus on the impact evaluation one year and then focus on the process evaluation questions in other years. See Table 5.2 for a recap of the enhancements made to the evaluation approach.

Data Collection, Instruments, and Systems. Whereas WR presumed that positive social-emotional changes for youths would affect their academic performance, measuring the indirect impact on academics was not the primary focus of their program. However, under NCLB, this perspective had to change. The challenge for the local evaluator was to create a feasible data-collection plan that incorporated these additional measures while not over-burdening the Woodcraft staff, school principals, teachers, or students. Methodologies used included participant satisfaction reports, youth surveys, parent interviews, and an outcome report for each participant completed by WR staff. WR's local evaluator believed that use of multiple methods, sources, and time points for triangulation of findings (Greene, Caracelli, & Graham, 1989) was both sensible and appropriate. The next section summarizes each of these methods and examines how each was revised to respond to the increasing emphasis on academic achievement outcomes.

Participation Records. Demographic (gender, ethnicity, age), school (current school and grade level), and program attendance data were

Table 5.2. Evaluation Approach Before and After No Child Left Behind

Initial Evaluation Approach	Adjustments Made to Evaluation Approach Influenced by NCLB
Focus placed on socioemotional outcomes while attempts were made to examine some academic outcomes	Increasingly focused on measuring academic outcomes, specifically standardized test scores, while maintaining some socioemotional measures
Paired sample *t*-tests conducted to measure changes for participants in socioemotional and academic outcomes	Added quasi-experimental design with comparison group regression analyses to assess the impact of the program on academic outcomes
Emphasis on evaluation for program improvement and meeting basic funder reporting requirements	Emphasis on impact evaluation while rotating inclusion of some process questions in different years

collected by program staff and stored in the Woodcraft Rangers tracking database. Even though Woodcraft had always collected this information, it was previously stored in Excel spreadsheets (or on hard copy) and children were not tracked by a unique identifier. With NCLB pressure to examine individual-level growth across time, the local evaluator adjusted the tracking system to address these emerging evaluation needs. Adjustments were made so that each student who attended the program obtained a primary key that uniquely identified him or her. This identifier also linked demographic information with student program attendance and responses on the Woodcraft-specific instruments (that is, youth surveys). In addition, Woodcraft staff were trained on how to collect the participants' unique school identifier, required to link them to their district data records.

Academic Records. WR experienced some benefits from NCLB related to accessing school data (discussed earlier), but there were other challenges in using standardized test scores as a measure of program outcomes. The primary challenge was changes in the tests used by the state. From 1998 to 2002, a norm-referenced test (the Stanford 9) was used in California, allowing local evaluators to conduct longitudinal analyses of individual student improvement. However, in 2003, this test was replaced by another norm-referenced test (the California Achievement Test 6th edition, or CAT6). These tests assessed how well each student was achieving academically compared to a national sample of students tested in the same grade at the same time of the school year. One year later, the state made another change and no longer required that the CAT6 be administered to all grade levels. Thus, the evaluation had to rely on one criterion-referenced test: the California Standardized Test (CST). The changes in tests made it impossible to use standardized test scores in longitudinal analyses (one of the goals under NCLB), given that individual student scores could not be compared across years.

The CST scores presented another challenge. The CST test was designed to be grade- and subject-specific to determine student proficiency levels: far below basic, below basic, basic, proficient, and advanced. This test was designed for cohort analyses, not for individual-level student tracking; therefore, it was considered inappropriate for longitudinal analyses of student progress. Evaluators are currently collaborating with other local evaluators and district evaluators to identify how best to use this test for purposes of evaluating an after-school program.

Participant Surveys. Participant surveys were designed prior to NCLB to measure aspects of youth development, the primary focus of WR. The evaluators, in collaboration with program staff, designed a two-page survey to be completed by students when they first joined the Woodcraft program and again at the end of each academic school year. The surveys assessed constructs related to school attitude, academic skills, personal sense of efficacy, problem-solving skills, and risk-taking behaviors. WR and their

evaluators made it a priority to keep the participant surveys in the evaluation plan even with the increased academic focus. Accepting the reality that many stakeholders wanted to know if the program contributed to improved test scores, WR did not want to exclude how youth-development results directly related to their activities. This was particularly important considering the difficulty of an after-school program to demonstrate a significant impact on test scores, when the evidence suggests that schools with 6 hours per day of instruction found it difficult to increase test scores substantially (Kane, 2004). However, recently WR included additional youth-development skills that could lead to academic improvement, such as leadership, study skills, and exposure to new activities.

The Woodcraft participant survey used a pre-post methodology, though it was not the most cost-effective approach given the staff time needed for administering it twice a year. Another approach often used to measure change is a retrospective measure, which would allow participants to report the extent to which the program made an impact. Retrospective change tools would require one instead of two administration points, and some researchers have found them to be more valid in measuring awareness, knowledge, and attitudes than pre-post measures (Pratt, McGuigan, & Katzev, 2005). To explore the validity of using this methodology with WR, during one year, the post student survey was revised to include a few retrospective items that reflected the same concepts as several of the items included in the pre-post measures. The findings from the retrospective change measures and the changes calculated from the pre-post measures were highly correlated. The retrospective change items were also strongly correlated with staff assessments (recently taken out of the evaluation). However, for this methodology to meet the scientific standards of rigor under NCLB and SBR, the surveys would have needed to include a control group and random assignment, which was not possible for WR given resources and program design. Although the pre-post design completed by program participants does not meet these standards, the program and local evaluator found that some funders still preferred the pre-post design.

Focus Groups and School Staff Interviews. Both prior to and after NCLB, qualitative data were collected from parents and school administrators through interviews and focus groups. These collection efforts were designed primarily for internal purposes of process evaluation, which was encouraged for program improvement (Gomby & Larson, 1992). They were conducted at the end of the program year with parents of participants and with school principals, soliciting their perception of program benefits and their satisfaction with the program. This was helpful for program improvement because it identified potential concerns and helped strengthen ties with the school sites. Given limited funds for evaluation, Woodcraft staff conducted the staff interviews and parent focus groups rather than the evaluation staff. Woodcraft primarily reserved the external evaluation budget for test-score

analysis and comparison-group activities. The evaluator did, however, design the instruments and protocols, trained Woodcraft administrative staff on how to conduct them, and incorporated the findings into the final report. This permitted collection of process data within the larger outcome evaluation. Although the quality of the data may have been affected given that program staff conducted the focus groups, this was one of the compromises required to create an evaluation that was responsive to the program's information needs within an NCLB framework.

During a year where there was a process evaluation focus, the evaluators were able to conduct a series of focus groups with students. The focus groups were conducted to obtain rich feedback about successful and unsuccessful recruitment and retention strategies. Half of the focus groups were conducted with students with less than 3 months of participation, while the other half was students who been in the program for more than 6 months. The sessions yielded rich contextual information that WR used for program improvement. The evaluators were able to conduct these sessions because comparison-group analyses on academic performance were not carried out that year. Unfortunately, limited evaluation funds could not accommodate both. The adjustments made to the data collection efforts and instruments are summarized in Table 5.3.

The Value of Evaluation at All Levels. Woodcraft leadership realized very early in the program's development that strong evaluation was necessary for growth and improvement. A key challenge after NCLB was to meet all of the internal and increased external expectations while still operating on a limited evaluation budget. WR realized the only way to accomplish this was for program staff to be actively involved in evaluation activities. Evaluation was integrated into the organization in a number of ways: (1) the

Table 5.3. Data Collection, Instruments, and Systems

Initial Data Collection, Instruments, and Systems	Adjustments Made to Data Collection, Instruments, and Systems Influenced by NCLB
Primary data-collection activities were implemented to document participant attendance and demographic information	Data collection systems became more sophisticated, enabling evaluators to merge data from different sources, including district files; increased attempts to examine longitudinal impacts
Participant survey designed to be cost-effective and focused on youth development	Participant survey adjusted to include more items related to academics; survey now using pre-post methodology
Qualitative research conducted by program staff through parent focus groups and staff interviews	Responsibility for qualitative research increasingly placed on program staff while evaluators focused on academic analyses; when this was not the case, evaluators played greater role in qualitative research

agency CEO promoted evaluation throughout the organization, (2) the agency invested in internal data systems, (3) all staff were assigned an evaluation role and received relevant orientation, (4) the agency and evaluator designed the evaluation plan together, (5) the agency and evaluator communicated at least once monthly on issues and progress, (6) the evaluator refined processes each year in response to lessons learned and staff input, and (7) the evaluator met with field staff once a year to discuss results.

Though the external evaluator collaborated with WR in designing the general evaluation plan, many aspects of the evaluation were carried out by WR staff, as described in prior sections of this chapter. In addition, WR's internal data department entered individual-level information in an Access database that was designed with the assistance of their evaluator. These data were then exported to the evaluator, who cleaned and merged the data and then conducted the analyses. By sharing the data responsibilities in this way, external evaluators reduced the budget while allowing the program to have direct access to information needed daily, such as enrollment and attendance rate. District data were obtained by the program through collaborative relationships with school districts and with the technical support of the evaluator.

Woodcraft had always conducted staff training with a focus on staff's role working with children. Increasingly, the staff training materials included information on data collection and evaluation. For example, after NCLB, the training binder for new staff presented an overview of evaluation—why it was conducted, what tools were used, how they were administered, and how the information was used to improve the program. As the pressure to produce high-quality, rigorous evaluation increased each year, so did the need to increase the Woodcraft staff's understanding of evaluation, their role in administering student surveys, and knowing how to interpret evaluation findings. Each year, the evaluators facilitated a meeting with the site coordinators to review the evaluation report and key findings. The intention was to gain staff buy-in for their role in the evaluation and create an opportunity to talk in more depth about the evaluation for program improvement. The meetings also enabled staff to examine results of academic impact that could be shared with principals to strengthen their relationship with the schools. A one-page summary of the report was created that each site coordinator could share with principals, placing an emphasis on academic outcomes. Woodcraft also started hosting an annual breakfast with principals from all participating sites to thank them for their collaboration. A summary of the latest evaluation was presented to principals, and they were engaged in discussion regarding lessons learned, challenges, and best practices that could be shared across school sites. These shifts in roles and responsibilities are listed in Table 5.4.

Program Improvement. Evaluation reports were designed to furnish detailed findings while also highlighting how to use the data for program improvement. The findings indicated academic outcomes were not attained

Table 5.4. Division of Evaluation Responsibilities

Division of Evaluation Responsibilities Between Program Staff and Evaluator, Pre-NCLB	Adjustments Made to Division of Evaluation Responsibilities Between Program Staff and Evaluator Influenced by NCLB
Minimal evaluation conducted so roles for program staff were limited	Program staff and evaluators shared roles in evaluation activities and staff training
Evaluation results primarily shared with funders and WR leadership	Evaluator also shared findings with field staff and field staff shared findings with school principals

without a high retention rate. From these findings and the pressure placed on school-linked programs to do more direct academic work, Woodcraft made programmatic adjustments, including using more formal curricula for the clubs and strategies for participant retention. One of the general changes observed since NCLB legislation was a shift in the type of out-of-school-time programs offered on school campuses. A recent study by the Center on Education Policy (2005) described a narrowing of the school curriculum to increase instruction time in math and reading by decreasing time spent on other subjects and enrichment activities. Though not as dramatic, a similar trend was observed in after-school programs. Although WR did not change its primary focus, it did incorporate formal curriculum into their clubs. The

Table 5.5. Program Improvement Efforts

Evaluation Findings	Resultant Programmatic Changes
Participants who attended more often had higher academic scores than those who participated less often *combined with* Participants were not consistently significantly different from nonparticipants in all of these areas	• Researched and purchased curriculum and trained staff on its use to increase integration of after-school math and literacy materials into the program • Added new positions—activities consultants— to provide ongoing expertise, coaching, and support in curriculum development for club leaders
Students who participated for longer periods of time improved more than those who participated less often *combined with* Students' average length of participation is 5–6 months	• Added incentives to increase student retention: – Created an incentive program based on program attendance – Built-in program enhancements of interest to youth (competitions, traveling specialists)
Middle school participants do not stay in WR within a given year as long as elementary participants	• Began MS youth retreats • Started structured incentive program • Expanded and formalized sports leagues and competitions • Restructured MS program so fewer sites assigned to each manager

curriculum promoted literacy and math skills through a "disguised learning" approach to increase the impact on academic performance. Table 5.5 gives several examples of how the program used outcome data for process purposes.

Conclusion

This focus on long-term academic outcomes might be referred to as a climate change that has affected not only schools directly under NCLB, but also, programs that are school-linked. The tenor of the debate in the American Evaluation Association regarding randomized control trials has certainly had a wide impact on evaluation within educational community-based programs. It has fueled the foundation world's increasingly strong focus on outcome-based measures and increased the pressure on evaluation practitioners to educate clients on long-term outcomes, particularly academic achievement. There has been an increase in the number of community-based organizations that feel the need to measure academic outcomes regardless of the nature of their programs. Thus, evaluators are often requested by clients to measure academic outcomes that are not linked to the goals of the program and for which there are few or no resources to adequately measure them. The WR case study constitutes an example of a program that increased its ability to address this pressure. The WR evaluation does not meet the SBR ideal, but the adjustments made have improved both the program and its evaluation.

Woodcraft is an example of how a community-based organization with school-linked programming has acknowledged, and responded to, changes in the school environment brought on by the NCLB legislation. Consequently, WR made appropriate programmatic enhancements to address the changing needs of schools and school administrators even though the program itself did not fall directly under NCLB. These enhancements have strengthened the program's relationship with schools and school districts. Woodcraft also had to make changes to its evaluation plan and instruments to demonstrate relevance to the schools, including how it can contribute specifically to improvement in student academic performance. These changes have been influenced by the NCLB evaluation guidelines and required close collaboration among the program, its local evaluator, and partner schools.

References

Center on Education Policy. (2005). *NCLB Policy Brief 3*. Washington, DC: Author. Retrieved March 28, 2007, from http://www.cepdc.org/document/docWindow.cfm?fuseaction=document.viewDocument&documentid=51&documentFormatId=686

Gomby, D. S., & Larson, C. S. (1992). Evaluation of school-linked services. In R. Behrman (Ed.), *The future of children: School linked services,* (pp. 68–84). Los Altos, CA: Center for the Future of Children and David and Lucile Packard Foundation.

Greene, J. C., Caracelli, V., & Graham, W. F. (1989). Toward a conceptual framework for multimethod evaluation designs. *Educational Evaluation and Policy Analysis, 11,* 255–274.

Kane, T. (2004). *The impact of after school programs: Interpreting the results of four recent evaluations* (working paper). New York: W. T. Grant Foundation.

Lodestar Management/Research. (2005). *Woodcraft Rangers annual evaluation report for 2003–04: Elementary and Middle School Programs.* Los Angeles: Author.

Lodestar Management/Research. (2006). *Woodcraft Rangers process evaluation report: Key factors related to program recruitment, retention and outcomes.* Los Angeles: Author.

Nichols, S., Glass, G., & Berliner, D. (2005). *High-stakes testing and student achievement: Problems for the No Child Left Behind Act.* Education Policy Studies Laboratory. Retrieved April 26, 2007, from http://epsl.asu.edu/epru/epru_2005_ Research_ Writing.htm

No Child Left Behind Act. (2001). Public Law No. 107–110. 107th Congress, 110 *Congressional Record* 1425, 115 Stat.

Pratt, C., McGuigan, W., & Katzev, A. (2005). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation, 21,* 341–349.

U.S. Department of Education. (2003). Notice of proposed priority: Scientifically based evaluation methods (RIN 1890-ZA00). *Federal Register, 68*(213), 62445–62447.

BLANCA FLOR GUILLÉN-WOODS is an independent consultant based in Washington, DC, and former senior research associate at Lodestar Management/Research.

MONICA A. KAISER is vice president of the Kaiser Group, a company that supports nonprofit organizations in achieving their mission.

MAURA J. HARRINGTON is the vice president of consulting services and senior researcher at Lodestar Management/Research.